THE UNIVERSITY
of EDINBURGH

Investigating transcription, replication and chromatin structure in determining common fragile site instability

Lora Boteva

THE UNIVERSITY
of EDINBURGH

Thesis presented for the degree of Doctor of Philosophy

The University of Edinburgh

2016

Declaration

I declare that except where specific reference is made to other sources, the work contained in this thesis is the original work of the author. It has been composed by myself and has not been submitted, in whole or in part, for any other degree, diploma, or other qualification.

Lora Boteva

September 2016

Acknowledgements

In the first place, I would like to thank my supervisor, Nick Gilbert, for his support and encouragment and for always finding the time for me throughout the three years. It has been both a privilege and a challenge to be supervised by him and I am very grateful for all the time, thought and efforts he put into my project and into training me as a scientist. I would also like to thank everyone else in the Gilbert lab, including Catherine, for convincing me to come to the lab and Ryu-suke, for his patience and for always setting a good example, as well as Covi, Adam, Sam, Peter, Hannah, Jim and Maria, who have been amazing to work with and learn from. I would also like to thank my family and friends, who provided much needed emotional support, in particular my mother and my brother, as well as my grandparents, who always gave me wise advice.

Abstract

Common fragile sites are a set of genomic locations with a propensity to form lesions, breaks and gaps on mitotic chromosomes upon induction of replication stress. While the exact reasons for their fragility are unknown, CFS display instability in a cell-type specific manner, suggesting a substantial contribution from an epigenetic component. CFSs also overlap with sites of increased breakage and deletions in tumour cells, as well as evolutionary breakpoints, implying that their features shape genome stability in vivo. Previously, factors such as delays in replication timing, low origin density and transcription of long genes have been implicated in instability at CFS locations but comprehensive molecular studies are lacking. Chromatin structure, an important factor that fits the profile of cell-type specific contributor, has also not been investigated yet.

Throughout their efforts to determine the factors that lead to the appearance of CFS lesions, investigators have focused on a single component at a time, potentially missing out complex interactions between cellular processes that could underlie fragility. Additional difficulties come from the cell-type specificity of CFS breakage: it indicates that only cell type-matched data would be informative, limiting the scope for studies using publicly available data.

To perform a comprehensive study defining the role of different factors in determining CFS fragility, I explored replication timing, transcriptional landscapes and chromatin environment across a number of CFSs in two cell types exhibiting differential CFS breakage. Initially, I characterised the patterns of CFS fragility in the two cell types on both the cytogenetic and the molecular level. I then used a FISH-based technique to investigate the process of mitotic compaction at active CFS sites and found that the cytogenetically fragile core of these sites sits within larger regions which display a tendency to mis-fold in mitosis. The aberrant compaction of these regions could be observed on cytogenetically normal metaphase

chromosomes, suggesting that finer scale abnormalities in chromosome structure underlie the cytogenetically visible breaks at fragile sites. I also investigated the links between transcription of long genes and CFS fragility using two approaches: I quantified levels of expression across all fragile sites using RNA-seq and modified transcription at a single active CFS using the CRISPR genome engineering methodology. My results indicate a complex interplay between transcription and CFS fragility: no simple linear correlation can be observed, but an increase of transcriptional levels at the active CFS led to a corresponding increase in fragility. To investigate the influence of the cell type specific replication programme and replication stress on CFS instability, I mapped replication timing genome-wide in unperturbed cells and under conditions of replication stress in both cell types. I found that replication stress induces bi-directional changes in replication timing throughout the genome as well as at CFS regions. Surprisingly, the genomic regions showing the most extreme replication timing alterations under replication stress do not overlap with CFS, implying that CFS instability is not fully explained by replication delays as previously suggested. Instead, I observed a range of replication-stress induced timing changes across CFS regions: while some CFSs appear under-replicated, others display switches to both earlier and later replication as well as differential recruitment of both early and late origins, implying that dis-regulation of replication timing and origin firing, rather than simply delays, underlie the sensitivity to CFS regions to replication stress. Finally, I investigated large-scale chromatin states at two active CFSs throughout S phase and into G2, the cell cycle stages most relevant stage for CFS breakage. I found that changes in large-scale chromatin architecture accompany the replication timing shifts triggered by replication stress, raising the possibility that such alterations contribute to instability.

In conclusion, I assessed the influence of multiple relevant factors on CFS fragility. I found that bi-directional replication timing changes and alterations in interphase chromatin structure are likely to play a role, converging to promote mitotic folding

problems which ultimately result in the well-described cytogenetic lesions on metaphase chromosomes and genomic instability.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AFM - Atomic Force Microscopy
APH - Aphidicolin
BAC - Bacterial Artificial Chromosome
bp - Base Pairs
BrdU - Bromodeoxyuridine
CFS - Common Fragile Fite
ChiP - Chromatin Immunoprecipitation
Chr - Chromosome
CIN - Chromosomal Instability
DAPI - 4,6-diamino-phenylindole
DDR - DNA Damage Response
DMEM - Dulbecco's Modified Eagle Medium (DMEM)
DMSO - Dimethyl Sulfoxide
DNA - Deoxyribonucleic Acid
ECL - Enhanced Chemiluminescence
EDTA - Ethylenedinitrilotetraacetic acid
EdU - 5-Ethynyl-2'-deoxyuridine
FACS - Fluorescently Activated Cell Sorting
FCS - Foetal Calf Serum
FISH - Fluorescent *in situ* Hybridisation
FITC - Fluorescein Isothiocyanate
FPKM - Fragments per Kilobase per Million Reads
HMM - Hidden Markov Model
HRP - Horseradish Peroxidase
LB broth - Luria-Bertani Broth
l - litre
MAA - Methanol: acetic acid, 3:1 fixative
NGS - Next Generation Sequencing
NTS - Nick Translation Salts
PCC - Premature Chromosome Condensation
PCR - Polymerase Chain Reaction
PFA - Paraformaldehyde
PI - Propidium Iodide
PVDF - Polyvinylidene Difluoride
RNA - Ribonucleic acid
RNAi - RNA interference
RT-PCR - Reverse Transcription Polymerase Chain Reaction
SDS - Sodium Dodecyl Sulfate
SMC - Structural Maintenance of Chromosomes
SNP - Single Nucleotide Polymorphism
SSC - Saline Sodium Citrate
TBE - Tris/borate/EDTA
TE - Tris/EDTA
TSA - Trichostatin A
UV - Ultraviolet
v/v - volume/volume
w/v - weight/volume

1. Chapter 1: Introduction

In mammalian cells, the long DNA molecules comprising the genome are wrapped in proteins to form a complex called chromatin. Chromatin fibres are then folded multiple times within the nucleus of the cell, allowing lengthy genomes to fit inside a much smaller nucleus. Apart from overcoming space constraints, the folding of the genome also has a regulatory function, influencing fundamental processes such as gene expression, genome replication and DNA damage repair (DDR) (Wolffe 1998).

As the physical context in which the genetic information is read, translated and maintained, chromatin plays an important role in preserving genome stability and responding to DNA damage; apart from serving as a structural template on which DNA repair takes place, chromatin components play an active role in the recruitment and retainment of a range of DNA repair factors. The role of chromatin in DNA repair has been studied extensively, but mostly in non-physiological systems where DNA breaks were induced by expression of exogenous restriction enzymes, lasers, UV and chemical mutagens (Kruhlak et al. 2006; Roukos et al. 2014). Although studies in these systems have generated valuable insights, a more current view of genomic instability is focused on the mechanisms through which internal factors and fundamental cellular processes such as transcription and replication also pose a risk to genome stability (Saponaro et al. 2014; Reijns et al. 2015). Unlike external factor-mediated instability, which usually arises from stoichiometric interactions of the damage-inducing agents with DNA and results in predictable outcomes, internally mediated instability is stochastic: it is likely to result from a combination of factors, including the exact chromatin context at the location where problems arise.

Common fragile sites (CFS), a set of genomic loci which become unstable in response to replication stress or oncogenesis, illustrate how a combination of

factors including transcription, replication timing and chromatin context are likely to play a role in inducing instability (Durkin & Glover 2007). As such, CFS represent a chance to study how the interaction of chromatin states and cellular processes may induce instability in a physiological system with disease relevance.

1.1 Chromatin components

Chromatin is a complex structure with different levels of organisation. At the most basic level, chromatin is made up of nucleosomes- 147 base pairs of DNA wrapped around a protein octamer consisting of 8 histone proteins. Arrays of nucleosomes are then further folded to form a fibre measuring 30-nm in diameter; the 30-nm fibres are then arranged into larger-scale structures forming domains with differing structural and functional properties, which ultimately form chromosomes-the largest units of chromatin organisation (Wolffe 1998). AN overview of chromatin organisation is shown in Figure 1-1.

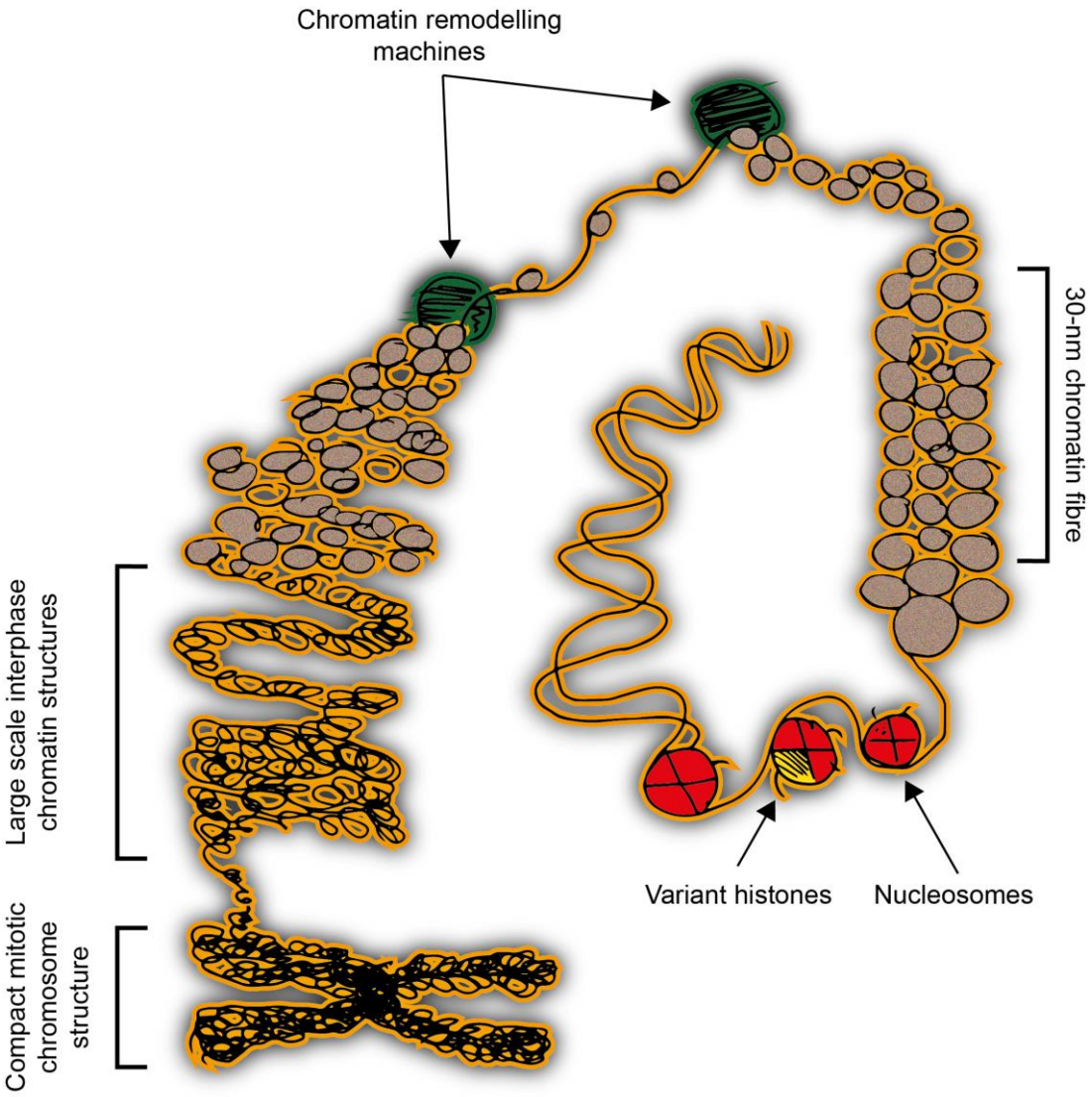


Figure 1-1 Overview of chromatin organisation. Multiple levels of chromatin organisation are depicted. At the primary level, the DNA strands are wrapped around nucleosomes, made up of classical and variant histones; variant histones shown in yellow (discussed in Chapter 1.1.1 and 1.1.2). Nucleosomal interactions give rise to the 30nm fibre conformation, which is disrupted in places to facilitate nuclear processes such as transcription and replication (Chapter 1.2.1); chromatin remodellers, shown in green, shift nucleosomes and alter their composition to induce alteration in the fibre structure (Chapter 1.1.4). Interactions between fibres give rise to large-scale structures, which are additionally folded in metaphase (Chapter 1.2.2 and 1.3.2). Figure reproduced from Boteva (2016).

1.1.1 Histones

Apart from DNA, chromatin contains numerous proteins with structural and regulatory functions. Among them, histone proteins are the most prominent. Core histones form nucleosomes-the basic repeat unit of chromatin, while linker histones form the connections between nucleosomes. Histone proteins can be post-translationally modified on their N-terminal tails, with different modifications exerting different effects on the chromatin fibre structure, adding a regulatory as well as a structural role to the range of histone functions. These post-translational modifications include acetylation, methylation and phosphorylation as well as other, less well-characterised marks. In addition to the canonical histone proteins the histone family also includes many histone variants, which can replace their classical counterparts in chromatin in a carefully regulated manner and in specific circumstances.

The histone proteins that form the nucleosome particle are called “core histones” and include H2A, H2B, H3 and H4 as well as their variants. Each nucleosome is an octamer consisting of two copies of each H2A, H2B, H3 and H4, arranged as an H3/H4 tetramer and two H2A/H2B dimers. The core histones are positively charged proteins, rich in lysine and arginine residues. They bind to DNA non-covalently, through electrostatic interactions between positive charges on histones and the negatively charged DNA molecule. Nucleosomes are separated by linker DNA, whose length is not constant, but can vary from 10 to 100 bp between species, cell types and genomic position along the DNA. Histone proteins binding to this linker DNA are called linker histones and include H1 and its variant H5, which is found in chicken erythrocytes (Harshman et al. 2013). Like core histones, H1 is also positively charged and is associated with both the linker DNA and the nucleosome particle. The H1 molecule consists of a globular domain and two tails, with the globular domain sitting at the nucleosome dyad, while the tails contact the linker DNA and drape along the chromatin fibre where it stabilises the folding of nucleosomes into a 30-nm fibre structure (Allan et al. 1980).

1.1.2 Histone variants

Histone variants-proteins with a high degree of sequence similarity to their canonical counterparts, can replace the classical histone molecules in the nucleosomes and linker regions; H2A, H2B, H3 and H1 all have non-canonical variants. Different variants replace canonical histones in the fibre in different circumstances in a replication-dependant, replication independent or tissue-specific manner. A class of proteins called histone chaperones carefully regulates this process. The incorporation of histone variants can have many effects, including a change in the fibre conformation (causing chromatin to become more or less tightly folded) or recruitment of regulatory proteins. Consequently, histone variants are indispensable for many biological processes, ranging from transcriptional activation to DNA repair and chromosome segregation (Henikoff et al. 2004).

An interesting example of a histone variant with a proposed dual structural and regulatory role is CENP-A, an H3 variant present specifically at centromeres, deposited by a histone chaperone called HJURP. The presence of CENP-A is necessary for recruitment of kinetochore components, but it is possible it also has a structural impact on the chromatin fibre. Although the precise effects of CENP-A incorporation into nucleosomes in vivo are unclear, a 2010 study found that nucleosomal arrays containing CENP-A are more condensed compared to arrays containing canonical H3, suggesting that the presence of CENP-A helps to establish an unusual chromatin structure at centromeres (Panchenko et al. 2011). Other well-characterised H3 variants include H3.1 and H3.3, which differ by just five residues in humans. H3.1 is incorporated into chromatin in a replication-dependent manner via the chaperone CAF-1 and is the major form of H3 in cells. Conversely, H3.3 can be integrated into nucleosomes at any time during the cell cycle via its own chaperone-HIRA; it is often found at transcribed genomic locations and promoters (Wirbelauer et al. 2005; Ahmad & Henikoff 2002).

Well studied H2A variants include macro H2A, found at the inactive X and H2AZ, a histone associated with active transcription which may promote nucleosome

destabilisation upon incorporation (Suto et al. 2000). As an important function of chromatin, locating and signalling DNA damage is associated with a separate histone variant-H2AX, an H2A variant representing around 11% of H2A in cells (West & Bonner 1980). H2AX replaces H2A in nucleosomes interspersed throughout the genome and differs from H2A in its C-terminal sequence. Upon DNA damage, the serine 139 position of the C-terminal portion of H2AX becomes phosphorylated. This phosphorylation is one of the primary events at sites of DNA damage and plays an essential role for DNA damage signalling, detection and repair.

1.1.3 Histone modifications

In addition to histone variants, chromatin fibre structure and composition can also be affected by post-translational modifications (PTMs) present on the N-terminal tails of the histone proteins. These modifications include acetylation, methylation, phosphorylation and ubiquination and similarly to the presence of histone variants, can act by directly modifying chromatins structure or by recruiting regulatory factors recognising specific post-translational marks. Numerous post-translational marks exist and their effects, both individual and combinatorial, are still under investigation.

Acetylation of lysine residues in the N-terminal tails of H3 and H4 is a mark associated with active transcriptional states. Acetylation neutralises the charge of the lysine residue, which is expected to weaken histone-histone and histone-DNA interactions, resulting in opening up of the chromatin fibre. However, a careful in-vitro study performed on short arrays of nucleosomes reconstituted on a repetitive DNA sequence failed to demonstrate significant opening of the chromatin, suggesting the effects of acetylation may depend on the wider chromatin context (Neumann et al. 2009). H3 and H4 can be acetylated at numerous positions, including H3K9, H3K14, H3K18, H4K5, H4K8, H4K12 and H4K16. Acetylation marks on H3 and H4 are recognised by bromodomains, protein domains present on some transcriptional activators and chromatin remodellers. The acetylation mark is added on histone molecules by a class of enzymes called histone acetyltransferases (HATs)

and removed by histone deacetylases (HDACs). The importance of maintaining acetylation states for genome stability is illustrated by the finding that loss of HDAC function is associated with instability, including aneuploidy and lagging chromosomes (Dovey et al. 2013).

Methylation occurs on lysine and arginine residues. Up to three methyl groups can be added on lysines, while arginines can only be mono- or di-methylated. Unlike acetylation, methylation does not change the charge of the residue affected. Lysines are methylated by lysine methyltransferases (HKMTs), which are very specific and methylate particular residues only. Multiple HKMTs have been identified, all of which share a domain called SET (Su(var)3-9, Enhancer-of-zeste and Trithorax). Arginine residues are modified by arginine methyltransferases, also known as PRMTs, while removal of the methyl residues is catalysed by demethylases. A few classes of lysine demethylases exist, including lysine-specific demethylase 1 (LSD1), which can demethylate different lysine residues depending on different accessory proteins and the jumonji domain demethylases, which act on tri-methylated lysine residues. Methyl marks on H3 and H4 residues can be associated with active and inactive chromatin states. Examples include H3K9me3-a repressive mark which recruits the heterochromatin protein HP1 and H3K4me3- a mark present in actively transcribed regions. Similar to the case for histone acetylation, a direct relationship between appropriate methylation patterns and genome instability has been demonstrated via depletion of Suv39h, an H3K9 methyltransferase involved in establishing H3K9me3 at pericentromeric chromatin. Mice lacking Suv39h are prone to tumour formation, while embryonic fibroblasts derived from the animals had extremely unstable karyotypes (Peters et al. 2001). While the H3K9 methylation mark probably exerts its effects on genomic stability through maintaining the structural state of certain genomic regions, another methylation mark, H3K79me, has been implicated in the DNA damage response in a signalling manner. This mark is established by the DOT1 lysine methylase and is important for recruitment of 53BP1, a protein integral to the DDR to a break site. 53BP1 recruitment by H379Kme is through a Tudor domain in the 53BP1 protein, a

domain recognising methylated residues; however, it is unclear whether the mark is established in response to a DNA break or whether chromatin changes within the vicinity of a break cause the mark to be exposed and recognised by 53BP1 (Huyen et al. 2004).

Another important PTM, phosphorylation, can be added on serine, threonine and tyrosine residues. This modification is placed by kinases and removed by phosphorylases. Unlike acetylation and methylation, which are related to establishing chromatin domains with different properties, phosphorylation has a major role in cell cycle progression. The serine 10 position on H3 is phosphorylated genome-wide by the Aurora B kinase as the cells progress through late G2 and into mitosis (Wei et al. 1999) in a manner that is inter-dependant with other histone modifications, such as H3K9me (Rea et al. 2000). This modification is required for the mitotic condensation of chromosomes- a process in which chromosomes are strongly compacted to facilitate chromosome separation and minimise entanglements during cell division.

As discussed earlier, phosphorylation of the H2AX histone variant at the serine 139 position (phospho-H2AX or gammaH2AX) is the most widely studied DNA-damage associated histone modification. This position is rapidly phosphorylated in response to DNA damage and phosphorylation is dependent on the ATM, ATR and DNA-PK kinases. The gammaH2AX mark spreads in large, megabase-sized domains surrounding the break region (Rogakou et al. 1999) and is essential for the DNA damage signalling and response. GammaH2AX-containing chromatin then serves as a platform for recruiting additional repair components, including 53BP1 and BRCA1 (Misteli & Soutoglou 2009). Interestingly, studies in which the H2AX phosphorylation site was disrupted indicated that lack of H2AX phosphorylation does not preclude initial recruitment of repair factors to the site (NBS1, BRCA1 and 53BP1); however, it interferes with their retention, reinforcing the concept of the gammaH2AX domain as a platform for retaining the factors necessary for repair. Following repair, H2AX phosphorylation is reversed by phosphatase complexes

including PP2A and PP4 (Chowdhury et al. 2005) and through histone exchange mediated by the FACT complex (Heo et al. 2008). Mammalian cells lacking H2AX exhibit enhanced susceptibility to genomic instability and cancer (Celeste et al. 2003). Given the coordinated structural and signalling functions of other histone modifications, it is tempting to speculate that phosphorylation of H2AX may act through structural effects on the chromatin fibre as well as through signalling in the DDR cascades. However, no such structural effects have been convincingly demonstrated to date and while changes in chromatin compaction are known to occur as a consequence of damage, they have been shown to be independent of the presence of gammaH2AX (Kruhlak et al. 2006). Other histone marks which may have a role in the DNA damage response include H2A ubiquitinylation (Ui et al. 2015), H2B phosphorylation at the serine 14 position (Fernandez-Capetillo et al. 2004) and H3 threonine 45 phosphorylation (Lee et al. 2015).

While the establishment of histone marks in response to DNA damage is well characterised, a recent publication by the Misteli lab explored the opposite idea-can certain histone PTMs predispose genomic regions to instability? Surprisingly, the study found enrichment of H3K4me1 and H3K27ac and depletion of the repressive H3K9me3 mark in genes frequently involved in translocations when compared to genes with similar expression patterns and levels (Burman et al. 2015). To demonstrate that the correlation is causal, they tethered H3K4 methyltransferase and H3/H4 lysine acetylase to a Lac operon (LacO) array also carrying an artificially introduced unique restriction enzyme site. When the frequency of breaks was assessed the authors found an elevated rate in the presence of both the H3K4 methyltransferase and the H3/H4 lysine acetylase, leading them to speculate that the more open chromatin environment promoted by these enzymes leaves the genome more exposed to instability. Interestingly the H3K4me3 mark has also been associated with the introduction of double stranded DNA breaks during V(D)J recombination in lymphocytes (Stanlie et al. 2010).

As illustrated by the many enzymes capable of establishing and removing PTMs, histone marks are transient states and subject to turnover during cell cycle and differentiation. The propagation of PTMs during DNA replication and cell division has been recently studied via nascent chromatin capture (NCC) - a technique based on the isolation and mass spectrometry analysis of nascent chromatin strands (Alabert et al. 2014). Application of NCCs in Hela cells showed that there is no erasure of PTMs during replication and that old histones carrying PTMs are integrated equally in the two nascent chromatin fibres. Newly integrated histones are then modified to copy the modifications already present in the fibre and PTMs are fully restored within the next cell cycle, with the exception of the trimethyl marks on H3K9 and H3K27, which appear to be accumulated slower over a number of cell cycles.

1.1.4 Chromatin remodellers

Apart from histones, chromatin contains a range of other proteins with diverse roles. One of the most important classes of non-histone proteins in chromatin are chromatin remodellers: proteins that can reposition and remove nucleosomes or change their composition in an ATP-dependant manner. Consequently, they introduce small-scale alterations in the state of the chromatin fibre and the accessibility of the DNA template. Chromatin remodellers are required for many nuclear processes, including transcription, replication, cell cycle progression and of course, DNA repair. Numerous mammalian chromatin remodellers exist and they can be broadly divided into four families: SWI/SNF, ISWI, INO80 and CHD.

The SWI/SNF family of remodellers includes two main complexes, BAF and PBAF. Both complexes contain an ATPase subunit (either BRM or BRG1) and three main core subunits: BAF155, BAF170 and BAF47. BAF and PBAF also can contain many different accessory subunits, which generate functional diversity and tissue specificity. The many functions of these enzymes include both transcriptional activation and repression and they are also implicated in tissue differentiation. Genes encoding remodellers of the SWI/SNF family are frequently mutated in

cancer and components of the SWI/SNF complexes, BAF and PBAF, have been shown to localise to sites of DNA damage. PBAF subunit BAF180 has a role in silencing transcription at sites of DNA breaks (Kakarougkas et al. 2014), while BRG1, the ATPase subunit common to BAF and PBAF, is involved in sister chromatid decatenation at the G2/M boundary and its inhibition results in anaphase bridges and lagging chromosomes (Dykhuizen et al. 2013). Hinting at the wide range of roles these remodellers have, the PBAF complex was also found to promote sister chromatid cohesion, especially at centromeres, with chromosomal breaks and abnormalities following its inhibition (Brownlee et al. 2014).

ISWI remodelling complexes are characterised by the presence of the SNF2H or SNF2L ATPase subunits as well as other accessory subunits; the family includes the complexes ACF, CHRAC and NURF, among others. These remodelling complexes also have numerous roles, can either act to activate or repress transcription and are also involved in replicating heterochromatin (Corona & Tamkun 2004). ACF-1, a component of two ISWI-type complexes, ACF and CHRAC, was also found to bind at laser-induced DNA breaks, co-localising with gammaH2AX (Lan et al. 2010). Cells depleted of ACF-1 were very sensitive to DNA damage, and the authors determined that ACF-1 facilitates the binding of NHEJ protein Ku at double strand DNA breaks.

The CHD class of remodellers derive their name by possessing chromodomains, which can read methyl marks on histones. An example is the nucleosome remodelling and deacetylase complex (NuRD), which promotes nucleosome compaction in heterochromatin. The CHD4 subunit of the NuRD complex is phosphorylated by ATM in response to genome damage (Polo et al. 2010) and is rapidly recruited to sites of damage (Larsen et al. 2010). In the same studies the authors observed increased rates of genomic breaks in CHD4-depleted cells, suggesting not only that CHD4 is essential for repair but that its depletion might make chromatin more susceptible to breaks. In contrast, NuRD complexes containing an alternative CHD3 isoform were released from heterochromatin upon

treatment with ionizing radiation, promoting chromatin relaxation (Goodarzi et al. 2011).

INO80-type remodelling complexes are characterised by an insertion within their ATPase domains, which leads to incorporation of Rvb 1/2 helicases, mammalian homologues of bacterial proteins mediating strand exchange and recombination. These complexes are thought to be involved in the turnover of H2AZ and H2AX in nucleosomes. INO80 is also involved in replication and S-phase progression, with fork collapse, slower S-phase progression and H2AZ mis-incorporation shown in yeast and mammalian cells in the absence of the complex (Hur et al. 2010). Mammalian cells depleted of the INO80 remodeller also exhibit DNA repair problems, with homologous recombination specifically affected as INO80 seems to be involved with 5'-3' resection of DNA at breaks sites (Gospodinov et al. 2011). Depletion of p400, an INO80 component primarily involved in the incorporation of the H2AZ variant at transcriptionally active regions, also makes cells sensitive to DNA damage (Courilleau et al. 2012). P400 was also shown to incorporate H2A.Z at double-stranded breaks, contributing to opening up of the chromatin in the break region to allow access for repair proteins (Xu et al. 2012). Another INO80 subunit, TIP60, which acetylates H2A and H4, has been implicated in restoring the chromatin environment following DNA damage response by removing the phosphorylated H2AX from the affected regions (Kusch et al. 2004). Another role for TIP60 includes acetylation of histones in heterochromatic breaks to allow chromatin relaxation for repair.

Overall, the study of the role of chromatin remodellers and their roles in maintaining genome stability is a very active field of research complicated by the many functions of these enzymes. In addition, as chromatin remodellers tend to have a serious impact on gene expression, studies have to exclude indirect effects on genome instability due to altered gene expression. This underlies the need for better and more representative in-vitro chromatin models, which would enable the

study of direct structural effects of the remodellers separately from their other roles.

1.1.5 Topoisomerases

The family of topoisomerases encompasses a number of enzymes present in chromatin, which act to resolve DNA catenanes and relieve topological stress by introducing breaks in the DNA strands and passing the strands around each other. Two main types of topoisomerases exist: type I topoisomerases introduce single strand DNA breaks, while type II topoisomerases break both strands of the DNA helix. Type I topoisomerases can be further classified into type IA and IB; type IA enzymes work to relax negatively supercoiled DNA, while type IB enzymes work by introducing a nick and allowing the strands surrounding the breaks to rotate relatively to each other and can act both on positively and negatively supercoiled DNA. Type II enzymes are categorised into two distinct families- IIA and IIB, based on similarity of structure and work by catalysing the passage of one DNA duplex through another.

Many chromatin processes, including transcription, replication and chromosome segregation generate topological stress and necessitate the action of topoisomerases. During transcriptional elongation, RNA polymerases generate positive supercoils ahead and negative supercoils behind and Top1 is involved in relieving the resulting topological tension. A 2013 study found that inactivation of Top1 affects genes longer than 200 kb specifically, underlying a possible association between Top1 mutations and autism (King et al. 2013). Secondary effects of Top1 may be related to avoiding direct and topological conflicts between the processes of transcription and replication, as illustrated by the finding that depletion of Top1 in mammalian cells causes replication fork stalling via transcriptional inference (Tuduri et al. 2009).

Topological tension associated with replication is thought to be resolved by topoisomerase II (Top2) enzymes. In addition to relieving topological stress generated during replication these enzymes may also play a role at sites of

converging replication forks (Baxter & Diffley 2008)- it has been speculated that replisome rotation during replication causes catenation of the daughter DNA strands behind the replisome, which can only be resolved through Top2 cleavage. In addition, Top2 enzymes also play a role in chromosome condensation, both functionally and structurally, with topoisomerase IIα (Top2A) shown to be a major component of metaphase chromosomes, forming a structural axis along the length of the chromosomes (Gimenez-Abian et al. 1995). Inhibition of Top2 through RNAi or inhibitors in various cell types and species has been shown to impair chromosome condensation and segregation to varying degrees, resulting in different chromosome morphologies in different organisms and cell types (Carpenter 2004; Uemura et al. 1987; Adachi et al. 1991). Conditional depletion of Top2A in human cells resulted in many serious defects, including activation of a G2 cell cycle checkpoint, partial condensation of chromosomes, anaphase abnormalities such as lagging chromosomes and generation of polyploid cells, underlying the role of this enzyme in cell cycle progression and chromosome segregation (Curanovic et al. 2013). However, as the compaction of chromosomes for mitosis and the resulting structures of mitotic chromosomes are still not completely identified, the precise role of Top2A in chromosome compaction and segregation is still unclear.

An interesting characteristic of topoisomerases is their ability to introduce single-stranded or double-stranded DNA breaks in a controlled manner. A number of topoisomerase poisons work by trapping the enzyme on the DNA strand following break generation, preventing the catalytic cycle from completion and resulting in a permanent break. As a result, the presence of these inhibitors causes genome-wide DNA damage and subsequent checkpoint activation, preventing progression through the cell cycle. Exploiting this property of topoisomerases and their inhibitors has led to a number of topoisomerase poisons becoming used as important anti-cancer agents (Ashour et al. 2015).

1.1.6 Condensin and cohesin

Another important class of chromatin components are the structural maintenance of chromosomes (SMC) proteins, a group covering two highly conserved complexes-cohesin and condensin. The main role of the SMC proteins is in maintaining and modifying chromosomal structure in the context of gene expression, mitotic compaction and chromatid separation.

Each of these two related complexes is composed of two SMC subunits and a kleisin-type component along with additional, complex-specific, subunits but always maintaining a ring-shaped configuration. In both complexes, the two SMC units form anti-parallel coiled coils that fold back on each other forming two "arms", which contact the kleisin subunit. For the cohesin complex, EM images estimate a length of 65 nm length for the SMC arms, potentially allowing chromatin fibres to be encircled by the complex (Haering et al. 2002). Due to this conserved ring-shaped structure of both of these complexes, it is thought that their interaction with DNA is topological rather than sequence-based (Murayama & Uhlmann 2014). Cohesin contains SMC1 and SMC3 as well as the kleisin subunit RAD21 and the non-kleisin unit SA (also called STAG). In the case of condensin, two different complexes exist in mammalian cells: both share SMC2 and SMC4 as components; condensin I carries the kleisin unit CAP-H and the additional subunits CAP-D2 and CAP-G, while condensin II has the CAP-H2 kleisin component as well as CAP-D3 and CAP-G2.

While cohesin and condensin share many similarities, their main roles are in a sense contradictory: the main role of cohesin is to hold sister chromatids together following the process of replication, while condensin acts to compact chromosomes in preparation of mitosis and facilitate separation of the chromatids.

Cohesin is first loaded onto chromosomes in early G1, but only becomes truly "cohesive" in S-phase. The initial loading is supported by the cohesin loader NIPBL in an ATP-dependent process. During G1, cohesin is continually loaded onto chromosomes and unloaded in a process mediated by WAPL and PDS5. In S-phase, the binding of cohesin to chromatin is stabilised and the complex acts to bring

together the two sister chromatids following replication until its release from chromosomes in mitosis. This release happens in two stages: at prophase, the complex is released from chromosome arms, due to phosphorylation by CDK1 and Aurora B to allow separation of the sister chromatids, while a small fraction is retained around centromeres; then, at anaphase, the kleisin subunit of the residual cohesin is cleaved by separase, allowing for chromosomes to be pulled apart by microtubules. The cohesin complexes are then deacetylated by HDAC8 and recycled back onto chromosomes in G1.

Historically, the cell-cycle related role of cohesin in sister chromatid cohesion was recognised first and this was subsequently followed by characterisation of its role throughout interphase. In addition to keeping replicated chromatids together, cohesin also has a role in organising large-scale chromatin structure at a global level, demarking chromatin domains and mediating enhancer-promoter interactions with resulting effects on gene expression. The first indication of this additional role for cohesin came with the discovery that cohesin binding sites along chromosomes coincide with CTCF binding sites, another protein involved in establishing large-scale chromatin domains (Parelho et al. 2008). It is thought that this role is executed by cohesin establishing topological linkages, similar to the linkages keeping sister chromatids together, but on a single DNA strand. Yet another role for cohesin was established with the finding that it is recruited to DNA damage sites following laser irradiation, but only in the S and G2 phases of the cell-cycle (Kim et al. 2002). The cell cycle specificity of this DDR repair recruitment of cohesin suggests that it may function in the homologous recombination (HR) DNA repair pathway, working to keep damaged chromatids and repair template in close proximity.

As a protein with such essential and diverse roles, mutations in cohesin subunits or in the proteins responsible for loading it onto chromosomes lead to serious pathologies, termed cohesinopathies. The best-studied cohesinopathy is Cornelia de Lange Syndrome (OMIM 122470), caused by mutations in all of the cohesin subunits, as well as in its loader, NIPBL. CdLS is clinically characterised by cognitive

impairment, limb defects and delayed development amongst others. Interestingly, the main defect of cells carrying CdLS mutations appears to be related with transcription rather than chromosome cohesion (Liu et al. 2009) suggesting that the most pathologically relevant consequences of cohesin depletion stem from its role in transcription regulation. Somatic mutations of cohesin subunits have also been described in various types of cancers (Solomon et al. 2011) including colorectal cancer and leukaemia (Barber et al. 2008; Kon et al. 2013).

The primary function of condensins is to assist folding of chromosomes in preparation for mitosis. Although extensively studied, the process of mitotic compaction is still unresolved and as a result, the precise manner in which condensins fulfil their primary function is still unknown. However, the structure of the condensin complexes and their localisation in interphase and metaphase are well described. Condensin in vertebrate cells exists as two distinct complexes, condensin I and condensin II. Within the context of mitotic folding, the two complexes have different roles. Condensin I acts to compact chromosomes laterally, while condensin II acts to compact chromosomes longitudinally-surprisingly, the exact ratio of condensin I to condensin II was found to impact on chromosome shape: samples prepared with *Xenopus* extracts with high condensin I: condensin II ratios chromosomes were longer and thinner compared to chromosomes prepared in extracts with a 1:1 ratio (Shintomi & Hirano 2011). The two complexes show differential localisation for most of the cell cycle: condensin I-is excluded from nucleus during interphase and localises to chromatin following nuclear envelope break down in the early stages of mitosis; in contrast, condensin II shows nuclear localisation throughout interphase and remains associated with chromatin during mitosis (Ono 2004). However, it appears that condensin II changes its binding to chromatin during S phase: a 2013 study found that it is not stably attached to chromatin in G1 and early S-phase; instead, it becomes stably localised only to replicated regions in S phase, forming “sister axes” and also working to resolve sister chromatids (Ono et al. 2013). Interestingly, this process was mildly impaired in cells exposed to replication stress suggesting that the behaviour of condensin II in S

phase links successful replication to mitotic folding and sister chromatid resolution. Later in the cell cycle, as condensin I gains access to chromatin in prometaphase, the two condensin complexes localise in an alternating pattern to axial structures extending along the length of the chromatids. The condensins then dissociate from chromosomes in telophase as chromatin decondenses.

Illustrating the importance of the condensin proteins, their depletion in chicken DT40 cells has shown that these proteins are essential for survival and their absence results in polyploidy. Codensin II depletion results in a high incidence of anaphase bridges; condensin I depleted chromosomes are wider and shorter and show diffuse scaffolds, while condensin II depleted chromosomes do not show scaffold defects but appear to lack axial rigidity (Green et al. 2012).

1.1.7 Additional chromatin components

Beyond the components mentioned above, chromatin contains numerous other proteins with a variety of roles which reflect the tasks and challenges of transcribing and duplicating the genome while maintaining its integrity. An example is provided by a family of helicases-the RecQ helicases, named “caretakers of the genome”. Helicases are a class of ATP-dependent enzymes specialised in separating the two strands of a nucleic acid duplex. They are involved in replication, transcription and DNA repair. In humans, the RecQ family includes the Werner syndrome protein (WRN), Bloom syndrome protein (BLM), RECQL4, RECQL and RECQL5. These proteins share a conserved helicase domain as well as a common function in preserving genome integrity. The contexts in which the RecQs work to prevent instability have been particularly well studied in the case of WRN and BLM, which have replication-related roles and RECQL5, which works to avoid transcription-induced fragility.

The genome-protecting properties of BLM and WRN are evidenced by the fact that disorders caused by germline mutations of these helicases are characterised by an increased cancer risk and other marks of genome instability. Ex-vivo lymphocytes from patients suffering from Werner syndrome (OMIM 277700), a monogenic

disorder caused by mutations in the WRN protein, show chromosomal instability (Gebhart et al. 1988) and an increased sensitivity to replication stress (Pirzio et al. 2008). Cells derived from Bloom syndrome patients deficient in BLM exhibit increased rates of sister chromatid exchange (SCE) (Traverso et al. 2003). Both proteins change their localisation in response to replication stress and form foci upon hydroxyurea treatment (Constantinou et al. 2000; Bischof et al. 2001); they also share an ability to unwind x-shaped and forked DNA structures (Mohaghegh et al. 2001), associated with stalled replication forks and holiday junctions. Despite similarities between the two, it is thought that they have subtly different roles, as suggested by the different characteristics of Bloom and Werner syndrome and the dissimilar defects seen in cells derived from these patients. The consensus is that WRN is specialised in rescuing stalled replication forks while BLM resolves holiday junctions.

The genome care-taking role of another member of the RecQ family, RECQL5, has been revealed by focusing on the process of transcription (Saponaro et al. 2014). RECQL5 interacts with RNA Polymerase II and slows it down in regions that are difficult to transcribe. Long-term depletion of RECQL5 led to recurrent loss and gain of chromosomal segments, specifically at long genes. Depletion of RECQL5 has also been shown to increase cell sensitivity to the Top I inhibitor camptothecin (Hu et al. 2009) and live imaging of cells with laser-induced DNA breaks has demonstrated that RECQL5 is recruited to DNA damage sites (Popuri et al. 2012).

1.2 Chromatin structure

Investigating the fine details of chromatin structure is a challenging task. Chromatin forms a dense mass inside nuclei that cannot be resolved by current microscopy techniques. Observations of synthetic chromatin fibres by electron microscopy has been informative, but such systems leave out the complexity of living cells and provide very reductionist results. However, advancements in microscopy and molecular techniques for investigating chromatin structure have added to our understanding and hold further promise for the future.

1.2.1 Nucleosomes and the 30nm fibre

Independent of any variants or post-translational modifications that may be present, core histones are invariably arranged in nucleosome structures, containing two H2A:H2B dimers and two H3:H4 dimers. 147 bp of DNA are wrapped around each nucleosome, with 10-100 bp "linker" DNA, bound to the histone H1 linking up different nucleosomes. With the help of linker histones, arrays of nucleosomes are folded into a fibre measuring 30 nm in diameter, the exact structure of which is still under intense debate (Figure 1-1). A number of models have been proposed for the arrangement of nucleosomes in the 30 nm fibre structure, including a solenoid model, where nucleosomes are organised in a helical array, a zig-zag model with a zig-zag arrangement of nucleosomes and an irregular fibre model with irregular arrangement and spacing of nucleosomes. Various techniques have been used in the last three decades to try and resolve the structure of the 30 nm fibre, including variations of electron microscopy, X-ray diffraction and most-recently-super-resolution microscopy (Grigoryev & Woodcock 2012; Ricci et al. 2015). While successful observation of the 30 nm fibre structure is possible in chromatin reconstituted in situ and in some rare types of nuclei, it has proven impossible to resolve the fibres in vivo in nuclei, with chromatin appearing instead as a dense mass.

In reality, as chromatin structure is very dynamic in living cells, it is likely the structure of the 30 nm fibre in living nuclei is not homogenous but instead is made up of a mixture of the models proposed with some regions being more compact and others more disrupted. In another illustration of the structure/function rule, it has been shown that constitutively transcriptionally inactive parts of the genome show a regular folding at the 30 nm level while bulk genome has a less regular conformation interspersed with irregularities (Gilbert & Allan 2001). Nucleosomes can be moved and shuffled by chromatin remodellers to allow proteins such as transcription factors, replication-related proteins and DNA repair proteins to bind to the naked DNA template. It is easy to envisage how these movements of nucleosomes can introduce transient local disruptions in the chromatin fibre. A

frequently used method to investigate nucleosome occupancy and 30-nm fibre structure is performed by testing the accessibility of the naked DNA by nuclease digestion.

1.2.2 Large-scale structures

At a further level of chromatin organisation, interactions between the 30-nm fibres give rise to so-called higher order structures. The fine details of this level of organisation are unknown (although looping of fibres is likely to be involved) and currently not many methods are available to investigate the folding of higher order chromatin structures. Overall, these structures are organised into segments with differing functional properties, determined by a combination of sequence composition (AT:GC content) and the presence of different histone modifications and chromatin bound proteins. A simplistic and classical view is to split the genome into gene rich segments with more open structures and silenced regions containing repeats and satellites where the folding of higher orders structures are more compact. However, a more current classification of the differing properties of higher order domains splits them into five functional categories: yellow (constitutively transcriptionally active regions), red (tissue-specific active regions), blue (repressed development and differentiation-related regions), black (silenced regions containing genes) and green (constitutively inactive repeats and satellites) chromatin (van Steensel 2011). The first two categories contain the transcriptionally active portion of the genome, which is enriched in acetylated H3 and H4. The chromatin structure in such regions is likely to have more disruptions at the 30-nm level and more “open” structures at the higher order level, facilitating easy access of transcription, replication and DNA repair factors to the DNA template. In contrast, the chromatin structure within the other classes is likely to be more compacted and less dynamic. Processes that require access to the DNA template such as replication and DNA repair may necessitate chromatin remodelling to open up chromatin in these regions of the genome. In fact, some recent data suggests that permanently silenced regions may act as a barrier to the DNA damage response

and that breaks within these regions may take longer to detect and repair (Goodarzi et al. 2008).

1.2.3 Nuclear organisation of chromatin

Within cells chromatin is contained within the nucleus-a complex organelle shaping the 3D organisation of the genome. Positioning of the genome in the nucleus has important functional consequences; nuclear position is a significant characteristic of a locus, impacting on its transcriptional activity, replication timing and proximity to other loci. Changes in the nuclear positioning of loci accompany development and differentiation, demonstrating the biological relevance of nuclear organisation.

The exact positioning of loci within the nucleus is probabilistic-it is not the same in every cell but is guided by a set of rules. With few exceptions, in mammalian cells, gene-rich, transcriptionally active regions of the genome are located towards the nuclear interior, while the gene-poor and heterochromatic regions are located towards the periphery. As a result, rather than having precisely defined locations, chromosomes have preferred radial positions in the nucleus. Centromeres also tend to be located towards the periphery (Gilchrist et al. 2004), while telomeres are distributed through the nuclear volume.

The nuclear periphery is defined by its interaction with the nuclear lamina- a part of the inner nucleoplasmic membrane. The genomic regions that interact with the lamina are known as lamina-associated domains (LADs); they measure 0.1 to 10 Mb in size and overlap with chromatin features such as low gene density and repressive histone marks. LADs can be divided into a facultative and a constitutive class. Facultative LADs are cell type-specific, while constitutive LADs are shared between cell types. Interestingly, disruptions in the lamina structure have been associated with genome instability, as illustrated by a class of diseases known as laminopathies, caused by mutations in the genes coding for the proteins that make up the nuclear lamina. The best studied among them is the Hutchinson-Gilford progeria syndrome (HGPS), a rare premature aging syndrome caused by mutations in the *LMNA* gene. Cells from patients with HGPS show microscopically visible

disruptions to the shape of the nuclear envelope, loss of the heterochromatic protein HP1 at the nuclear periphery and altered histone modifications pattern. Although not deficient in any of the components of the DDR response, HGPS cells are sensitive to ionizing radiation and accumulate DNA damage when grown in culture (Musich & Zou 2011). They also display increased levels of γ-H2AX and ATR/ATM activation.

1.2.4 Chromosome territories

Rather than being dispersed throughout the nucleus, each chromosome occupies a distinct volume, called a chromosome territory. This has been demonstrated by chromosome painting-a FISH-based technique where the genome is hybridised to a large number of chromosome-specific probes to allow visualisation of individual chromosomes within the nucleus. The radial positioning of a chromosome is strongly influenced by its composition -gene-poor chromosomes tend to occupy positions closer to the nuclear periphery while gene rich chromosomes are more frequently located towards the interior (Boyle et al. 2001). This trend is illustrated by human chromosomes 18 and 19, which are very similar in size but have very different sequence composition: chromosome 18 is gene poor, while 19 is gene-rich. The Bickmore lab used chromosome territory FISH to investigate the positions of the two chromosomes in the nucleus and found that chromosome 18 was consistently located closer to the nuclear periphery than chromosome 19 in both lymphoblastoid and fibroblast cell lines (Croft et al. 1999). The radial positioning of chromosomes in the nucleus was also found to be tissue-specific, with more closely related cell types exhibiting more similar chromosome positioning (Parada et al. 2004). The human genome also contains 5 acrocentric chromosomes, containing rDNA sequences –chromosomes 13, 14, 15, 21 and 22 which are usually clustered around the nucleolus-the site of transcription and processing of ribosomal RNA.

The radial rule of chromosome positioning also influences the positioning of alternating gene rich and gene-poor segments within chromosomes-in this case, gene rich segments are located more centrally while gene-poor regions occupy

regions closer to the periphery. In addition, within chromosome territories, transcriptionally inactive segments are located internally and transcriptionally active segments are at the surface of the territory (Boyle et al. 2011). This arrangement allows transcriptionally active regions ready access to the transcription machinery and domains rich in mRNA metabolic factors such as SC-35 foci (Shopland et al. 2003). However, the fine-detail structure of chromosome territories is yet unclear, reflecting our lack of knowledge of the chromatin structures that shape them.

From a genome stability perspective, an important consequence of chromosome positioning patterns relates to translocations, the most frequent chromosomal abnormality seen within the human population. It is well established that the physical proximity of two chromosomes in the nucleus affects the probability of a translocation occurring between them. An analysis between the frequencies of different non-pathogenic translocations in the human population and the preferred radial positions of chromosomes in the nucleus found that chromosomes with similar nuclear positions form translocations more frequently than expected by chance (Bickmore & Teague 2002). Another study was able to demonstrate close proximity between the BCR and ABL loci, involved in the well characterised t (9; 22) translocation forming a “Philadelphia” chromosome in chronic myeloid leukaemia. The authors showed that the BCR and ABL loci were closer in B-lymphocytes than in hematopoietic progenitor cells, suggesting that cell-type specific aspects of nuclear organisation may contribute to the association of certain translocations with particular cancer types. In 2013, the Misteli lab published a study exploring the dynamics of double strand breaks and subsequent translocation formation in an elegant system: NIH3T3duo cells encode a small number of *Sce*I restriction enzyme sites integrated on different chromosomes, with some sites adjacent to a *LacO* array, while other sites neighboured a *TetO* array (Roukos et al. 2013). Upon break induction by the *Sce*I enzyme, it was possible to track the breaks which were marked by fluorescently tagged Lac (LacR) and Tet (TetR) repressor proteins; translocation formation was indicated by long-lasting, stable co-localisation of the

LacR and TetR signals. The authors were able to demonstrate that most translocations are formed by loci that are closely located prior to break induction (contact-first model) rather than as a result of a movement of double strand breaks to proximal locations (breakage-first model).

1.2.5 Nuclear scaffolds

An unresolved question, widely discussed in the 1980s and reframed today is whether chromatin and other nuclear components are freely suspended in the nucleus or are instead attached to an underlying nuclear structure. Careful observations by electron microscopy found some evidence of an internal nuclear matrix, composed of irregular fibres connected to the nuclear lamina. These structures were resistant to DNase treatment but vulnerable to RNases, which suggested the presence of an RNA component. Various biochemical approaches were used in attempts to characterise the exact components of these nuclear structures and their relationships to the genomic sequence. Techniques aiming to isolate the genomic sequences associated with the matrix showed that satellites and G-bands were over-represented (Craig et al. 1997) . The search for the protein components was less successful: a family of ribonucleoproteins, hnRNPs, were thought to be involved but due to an inconsistency of experimental procedures, the full composition and structure of the nuclear matrix was never resolved and the idea was abandoned. Although the concept as defined in the 1980s is unpopular today, recent literature has also addressed the role of non-coding RNAs as possible molecular scaffolds. A 2014 study found that C_0T-1 RNA, transcribed from interspersed repetitive elements, associates with chromatin in a fractionation resistant manner, occupies defined chromosome territories and persists after transcriptional inhibition (Hall et al. 2014). A well-studied example is also provided by XIST, a 19 kb long non coding RNA expressed from the X chromosome, which is essential for the process of X chromosome inactivation. XIST localises precisely to the territory occupied by the inactive X and maintains its association even after histone extraction (Clemson et al. 1996). Yet another case is HOTAIR, a long non-coding RNA which associates with target genes and targets the silencing PRC2

complex (Rinn et al. 2007), illustrating the duality of functional and structural roles integral to most chromatin components.

One of the hnRNP proteins identified as a component of the nuclear matrix is hnRNPU, also called SAF-A. SAF-A is an abundant nuclear protein which was first identified among proteins bound to the matrix attachment regions (Romig et al. 1992). It has been implicated in many processes, including splicing, mitotic progression and DNA repair (Britton et al. 2014; Ma et al. 2011). Recent data from the Gilbert lab (Ryu-suke Nozawa, mnuscript in preparation) indicates that SAF-A affects DNA structure at transcriptionally active regions- depletion of SAF-A results in a change towards more compacted chromatin states. Interestingly, SAF-A depletion also resulted in an increase in gammaH2AX staining, implying an exciting link between this structural chromatin component, chromatin compaction and genome instability.

1.2.6 Methods for investigating large scale chromatin structure

Two complementary methods are often used to study the 3D organisation of the genome at the level of higher order domain structure: FISH-based methods and chromosome confirmation capture methods (Bickmore & van Steensel 2013). FISH relies on hybridisation of fluorescently labelled probes to visualise individual loci, defined portions of the genome or whole chromosomes. It provides a snapshot of nuclear structure at the single cell level, but disadvantages are that it is time-consuming and provides a limited amount of information at a low resolution. Chromatin conformation capture (3C) techniques rely on “freezing” the nuclear structure by cross-linking interactions within the nucleus, ligating DNA fragments held in proximity by the cross-links, followed by PCR or next-generation sequencing to identify hybrid DNA fragments, indicative of contacts. At the most sophisticated end, these techniques can theoretically identify all possible interactions throughout the genome, but there are also disadvantages. Unlike FISH, 3C techniques work on populations of cells rather than at a single cell level, producing a population average

which may reflect a number of different contact configurations at the single cell level. Despite the caveats, 3C methodologies have been very influential in the field of 3D genome organisation, contributing the concept of topologically associating domains (TADs), defined as regions measuring ~900 kb, where contact maps show increased interactions. The full human genome is divided into approximately 2000 TADs which also overlap with the distribution of histone marks and other genomic features such as replication timing (described in Section 1.2.7). However, they are not cell-type specific and the question of what level of structural organisation they reflect and their functional importance is still open to debate. Interestingly, the translocation frequency pattern seen with chromosome territories can be also traced to the TAD level of organisation-a study conducted in B-cells found that the likelihood of translocation between two loci is strongly related to the contact frequency between them, as defined by chromosome confirmation capture-generated contact maps (Zhang et al. 2012).

As the only techniques currently available to study higher order chromatin structure, a natural question is how to reconcile the results produced by conformation capture techniques and FISH. FISH-based studies have shown that probes located within a TAD are physically closer than probes not located within the same TAD but separated by a similar “linear” genomic distance (Bickmore & van Steensel 2013; Nora et al. 2012). On the other hand, confirmation capture generated maps sometimes indicate physical contacts between loci separated by very large linear distances, such as whole chromosome. A study at the HoxD locus drew attention to the fact that results from the two techniques are not always compatible and urged caution when interpreting results (Williamson et al. 2014).

Despite caveats, FISH and conformation capture techniques have transformed our understanding of higher order chromatin structure and many efforts are under way to improve them and transcend their limitations. The chromosome conformation capture technique has been combined with sequence capture methodology to yield Capture-C, a technique that allows higher resolution mapping at defined genomic

loci for reduced sequencing costs. A high-throughput FISH assay has also been developed (Shachar et al. 2015). An exciting development has come from utilising the Cas9-CRISPR system to label and monitor endogenous loci using live imaging (Chen et al. 2013). As a novel method, this technique has been mostly limited to repetitive loci and currently requires a large number of gRNAs which render it impractical for most uses. However, it promises to provide unprecedented insight into real time chromatin dynamics in future.

1.2.7 Transcription and replication in the nucleus

The structurally and functionally complex chromatin structure described above exists as the background for the two most important genomic processes- transcription and replication. Therefore, chromatin should always be considered in the context of these two fundamental processes. As we have seen above, the nucleus is a site of many correlations: radial position, gene density, histone mark enrichments and transcriptional activity.

Another correlation comes from the process of replication: the exact timing of replication of a locus also correlates with its nuclear position, as well as with its transcriptional activity. Replication proceeds in a well-controlled timely manner across the genome: alternating segments of chromosomes replicate at different times throughout S-phase, with gene-rich, transcriptionally active segments replicating early in S-phase and heterochromatic regions replicating last. These replication domains measure from 400 to 800 kb and control of replication timing is achieved by simultaneous firing of clusters of origins within the replication domains at defined times during S-phase. The correlation between replication timing and nuclear position is so strong that it gives rise to striking S-phase patterns visible in nuclei stained with markers of active replication early replicating cells show diffuse staining with markers excluded from the nuclear periphery; cells in mid-S have speckled patterns; and in nuclei in the latest stages of replication the staining overlaps with the nuclear periphery and heterochromatic regions. Replication timing domains partially overlap with TADs, however some replication domains are

cell-type specific and change during development and differentiation, along with changes in transcription. About 80% of the genome has constant replication timing between cell types, with 50% showing development and differentiation-related changes (Ryba et al 2010).

A few studies to date have tried to separate out the effects of chromatin state, transcription and replication timing to investigate the real determinants of nuclear positioning. A recent study by the Bickmore lab indicated that the chromatin compaction state may be the primary factor, whilst replication timing was shown to be a consequence of transcriptional state (Therizols et al. 2014). However, other studies have argued that replication plays a role in the establishment of nuclear organisation. A recent chromatin confirmation capture study revealed that TAD structure is established during early G1, at the same time as the replication timing program (Dileep et al. 2015). Another recent study used high throughput FISH to screen for factors affecting nuclear positioning of a small number of loci; it found that a number of replication-related proteins significantly affected positioning and also that replication was needed to maintain correct nuclear positioning (Shachar et al. 2015).

1.3 Chromatin changes throughout the cell cycle

Most of the in vivo studies in the field of chromatin research have been focused on cells in interphase and G1 cells in particular. However, chromatin structure is not static and undergoes many changes throughout the cell cycle to facilitate the replication of the genome and its equal separation between daughter cells during the process of cell division.

1.3.1 Chromatin changes during replication

The duplication of the genome is a challenging time for chromatin: it necessitates the disruption of chromatin fibres to individual nucleosomes, their subsequent reassembly, maturation of the nascent chromatin fibre and re-establishment of the preceding epigenetic states.

The replication fork passage through a chromatin fibre causes displacement of nucleosomes. The rate of nucleosome displacement is very high: assuming a fork speed of 2-3 kb/min amounts to 10-15 nucleosomes displaced per minute. Upon displacement, each parental nucleosome is split into two H2A-H2B dimers and an H3-H4 tetramer, which remain in the vicinity of the replication fork. These parental histones are then recycled back onto the nascent DNA strands, split in a random manner between the parental and the daughter. The recycling of parental histones is not an entirely passive process: histone chaperones, such as FACT and ASF1 act as acceptors of the ejected histones and deposit them onto the newly replicated DNA. These recycled parental histones carry the PTMs they have acquired prior to replication, allowing the local histone landscape to be propagated. In addition to the recycled histones from the previous cell cycle, newly synthesised histones are also deposited on the nascent DNA-experiments utilising SILAC pulses to mark old histones showed a nearly equal rations of old to new histones in replicated chromatin (Alabert et al. 2015). In the case of H3 and H4, they are placed as a dimer by the chaperone CAF-1, which in turn is targeted to sites of active replication via an interaction with the replication clamp protein, PCNA. A well-known characteristic of newly synthesised H3 and H4 histones is that they are acetylated; the acetylation marks are then removed 20 to 60 minutes following replication. An interesting suggestion is that the presence of these acetylated histones on the nascent chromatin allows a window of time in which the chromatin structure is more open and amenable to DNA repair processes and re-establishment of transcription. However a failure to remove that de-acetylation can be detrimental: HDAC1 can be found at sites of active replication and it is possible that some of the negative effects of HDAC inhibitors could be associated with a failure to restore the chromatin state following replication (Milutinovic et al. 2002). The demand for synthesis of classical histones is unique to the S-phase of the cell cycle. Their production is tightly regulated to avoid both insufficient histone supply and accumulation of unincorporated histones. The rate of replication is also affected by the availability of histones; DNA combing experiments in conditions of reduced

histone supply showed that fork speed slows down and S-phase was prolonged (Mejlvang et al. 2014).

Following the initial deposition of histones, chromatin “matures”. This process can be assayed by digestion of newly synthesised chromatin with DNase. Such experiments were first performed as early as the 1980s and found that recently replicated DNA was more easily digested than bulk chromatin (Cusick et al. 1983). Later studies demonstrated that newly synthesised chromatin starts to produce a nucleosomal periodicity comparable to pre-replication chromatin within 10 to 20 minutes post-replication. This reinstatement of the preceding chromatin structure is termed “maturation”.

Further to the question of restoring chromatin structure post replication is the question of restoring epigenetic states. The maintenance of epigenetic states in chromatin is a complex process, impacted by many modifying enzymes and transcriptional activity. During replication, these states are transiently disturbed; in most cases, epigenetic states are restored through the incorporation of parental histones and the symmetrical recovery of the marks they carry. However, this is not the case with states determined by some histone variants which cannot be incorporated during replication- for example, domains of H2A.Z marking transcriptionally active regions are depleted after replication. Comparison of histone composition of pre and post-replication chromatin shows that histone marks are diluted two fold in newly assembled chromatin and are then gradually restored until the marks reach pre-replication levels, contrary to an alternative model where marks are restored immediately on new histones in replication-coupled manner. An extreme case are the H3 repressive marks H3K9me3 and H3K27me3- full restoration for them can take more than a single cell cycle (Alabert et al. 2015). Interestingly, replication stress has been shown to affect histone marks on ASF1 associated histones, suggesting it may interfere with the re-establishment of epigenetic states (Jasencakova et al. 2010).

1.3.2 Chromosome compaction in mitosis

Following replication, the next major cell-cycle related change is the compaction of chromosomes for mitosis. The folding of chromosomes in preparation for mitosis is the most radical structural change the genome undergoes throughout a cell's lifetime. It achieves a 10,000 times level of folding compared to the linear DNA molecule and results in reproducible chromosome morphology and binding patterns. Despite the fact that this process has fascinated biologists for decades, neither the underlying molecular events, nor the resulting fine structure of the fully compacted metaphase chromosomes are resolved at present.

Attempts to deduce the chromatin structure of a fully compacted metaphase chromosome can be traced back to the 1970s, when SEM and TEM techniques were first used to image isolated chromosomes. Decades of electron microscopy imaging and multiple adjustments to the sample preparation methods yielded many models of chromosome folding, some of which show significant compatibility with modern molecular data. Among them are the radial loop model and the successive helical coiling model. In the radial loop model, loops of higher order chromatin fibres are arranged radially around a proteinaceous scaffold. The helical coiling model is based on observations of metaphase chromosomes prepared under specific conditions (such as hypotonic shock and in the presence of urea) in which the sister chromatids are seen to follow a helical, zig-zag path. Similarly to the radial loop model, chromatin loops are radially arranged around a proteinaceous scaffold, however in this model the scaffold is twisted in a helical path. More complicated models envision a hierarchy of folding structures which may be not be sequentially formed (Belmont et al. 1987). In recent years, novel imaging approaches have signalled the end of the EM era of chromosome analysis. Nowadays, atomic force microscopy (AFM) and super-resolution methods such as SIM and STORM promise to provide further insights into mitotic chromosome structure, however passive observation of isolated chromosomes will always have a supplementary role to molecular investigations of the processes involved.

Atomic force microscopy (AFM) is a type of microscopy which uses a scanning probe with a sharp tip over the surface of a sample, measuring the force between the probe and a sample to determine the properties of the sample. AFM studies on metaphase chromosomes have identified features named ridges and grooves along the length of the chromatids. Ridges correspond to the Giemsa positive, AT-rich bands (G-bands) and grooves correspond to the Giemsa-negative R-bands. Ridges and grooves are roughly symmetrical on sister chromatids. In addition to the ridges and grooves, globular structures can be seen on surface of chromosomes, which appear to be produced by strongly twisted fibrous structures measuring 50-nm in size. The twisted structures are more tightly packed in ridges and sparser in groves. This observation could reflect an underlying structural difference between the folding of gene-poor and gene-rich bands or could reflect preferential extraction of proteins from the grooves during the preparation procedure. Another observation from AFM imaging of mitotic chromosome is the presence of fibrous structures measuring 50-60 nm which appear to connect the two sister chromatids. These structures are present throughout the chromosome in prometaphase, but can only be seen in ridges in late metaphase (Ushiki & Hoshi 2008).

A major drawback of EM and AFM is the complex sample fixation and preparation procedure, which may distort the chromosome structure and limit information about the behaviour of chromatin in live cells. Due to that, live imaging studies may be needed to complete the picture of mitotic folding. A good example is a 2015 study using live imaging in cells expressing fluorescent H2B, allowing observation of chromatin movement in the period between prophase and metaphase (Liang et al. 2015). Authors found that chromatin became less dense from mid-prophase to late prophase, then denser in metaphase, suggesting that chromatin expands and then contracts in the run-up to metaphase. In addition, they could observe the individualisation of each sister chromatid in late prophase by imaging topoIIα, followed by a radial rearrangement of chromatin around the newly separated sister chromatid axis.

As a step away from imaging based methods, chromosome confirmation capture has also been used in an attempt to define the mitotic chromatin structure (Naumova et al. 2013). However, these experiments were more successful in stressing the differences between interphase and mitotic chromatin organisation than defining mitotic structures. The contact maps generated from cells in G1 and S phase differed sharply from contact maps from mitotic cells; the well-described TAD compartments present in interphase cells were absent in mitosis. Instead, a linear decrease in contact probability could be seen with an increasing genomic distance up to 10 Mb, followed by a sharp drop off in the contact frequencies for distances larger than 10 Mb, reflecting the linear nature of mitotic chromosomes. However, it is arguable if Hi-C, a technique optimised in interphase cells, can be informative when applied to the dense, protein-rich structures of mitotic chromosomes.

In addition to the histone: DNA fibres, metaphase chromosomes have a central structural axis made up of topoIIα and condensin I and II. The concept of chromosomal axis was first defined in experiments in which histones were removed from chromosomes; histone-depleted chromosomes appeared as halos of DNA, interpreted as chromatin loops, surrounding a central proteinaceous structure (Adolph et al. 1977). The scaffold is thought to anchor chromatin loops, however its fine structure and mode of interaction of DNA is not defined. TopIIα, condensin I and condensin II can be seen localising to the axis in an alternating manner and depletions of either one of the three proteins results in changes in chromosome morphology. In addition to changes in morphology, condensin depletion, also changes the elasticity of chromosomes, suggesting that axis formation is important for withstanding the stress of chromosome segregation and potentially, the contraction and expansion cycles leading to compaction. Live imaging has shown splitting of the axis during late prophase, which in turns precedes the parting of sister chromatids.

Another important signal of mitotic compaction is the cell-cycle dependent phosphorylation of the serine 10 position of H3. H3 phosphorylation, driven by the

Aurora B kinase, starts to appear in late G2 and peaks at metaphase. Good correlation can be observed between H3 phosphorylation and chromosome condensation and it is well-known that drugs which can induce H3 phosphorylation also induce premature chromosome condensation-a process in which chromosomes condense independently of cell cycle stage (Wei et al. 1999). However, such drugs are not specific enough to establish causality between H3 phosphorylation and condensation. H3 is de-phosphorylated in anaphase in a process which precedes microscopically visible decondensation in telophase. However, dephosphorylation of H3 has also been described in the absence of chromosome decondensation, suggesting a complicated relationship between the two (Magalska et al. 2014).

An interesting example of specialised mitotic chromosome structures is presented by telomeres. Telomeres are ribonucleoprotein structures composed of kilobases of conserved TTAGGG repeats, covered by the protective shelterin complex. Telomere structures mask the ends of chromosomes to prevent DNA damage signalling and inappropriate repair; instead, a single stranded, G-rich strand at the end of the chromosome invades the conserved repeat sequences, forming a structure known as a T-loop. A helicase, RTEL1, has been shown to disassemble T-loops to facilitate replication of telomeres (Vannier et al. 2012). Depletion of the shelterin component TRF1 led to activation of DNA damage signalling and the appearance of “fragile telomeres”, suggestive of disrupted chromatin structure at telomeres in mitosis (Sfeir et al. 2009).

1.3.3 Decompaction and reorganisation of nuclear architecture

The process of reversal of mitotic folding and restoring interphase chromatin structure is also not well understood. Broadly, two stages, likely driven by different processes, can be recognised. The cytologically visible compaction is lost in the transition between anaphase and telophase, while a full re-establishment of nuclear architecture is not achieved until two hours into G1, when chromosome territories assume their preferred positions and the topological compartments are restored.

The initial, cytologically defined stage of decompaction is extremely fast and the underlying molecular events are not well understood. It is debated whether this quick decondensation results from a simple reversal of the molecular changes implicated in the compaction process: for example, condensin release from chromosomes and de-phosphorylation of H3, both of which happen in anaphase. However, observations of the decondensation of purified mitotic chromatin in *Xenopus* egg extracts indicated that ATP and GTP were necessary for decondensation and identified RuvB ATPases as factors that actively participate in the decondensation process (Magalska et al. 2014).

Further to the microscopically obvious decompaction, finer-scale chromatin changes appear to be under way in early G1. As demonstrated by HiC experiments, the compartmentalised structure of the genome is lost in mitotic chromosomes. TAD domains and other features of nuclear organisation, such as the preference of gene-rich loci for the interior and gene-poor loci for the periphery are not re-established in the process of anaphase to telophase decompaction. Instead, these features are recovered around 2 hours into G1, coinciding with the replication-related origin decision point, in which the origins which will fire in the next S-phase are selected and the replication timing program is initiated (Dileep et al. 2015). The molecular determinants of these two concomitant processes are unknown, but they could be driven by the presence of epigenetic marks or resumption of transcription by "bookmarking" transcription factors. Interestingly, HiC experiments at defined time points post mitosis showed that progression towards the re-establishment of TAD compartmentalisation is continuous and not sudden however it is unclear if that is a real finding or a result of asynchrony in the cell population.

1.4 Chromatin and DNA repair

A cell's genome is frequently exposed to factors that have the potential to introduce changes in the DNA sequence ranging from point mutations to chromosome structural aberrations and even chromosome gain or loss. Classically, threats to genome integrity were perceived to come from external factors, such as

drugs, chemical compounds or UV radiation. A more current view is that internal factors and fundamental cellular processes such as transcription and replication also pose a risk to genome stability.

Whatever the source of the threat, chromatin is the context in which the genome is assaulted and then repaired. However, chromatin is more than just a passive background in the DNA damage response. It forms a dynamic structure that plays an active role in a cell's response to genome damage and reacts to DNA damage with extensive changes to its structure and composition. The best accepted model describing chromatin dynamics upon induction of damage is the so-called "access, repair, restore" model (Green & Almouzni 2002). It postulates that to fully repair a damaged locus, chromatin first has to be disrupted to allow access to the damaged template, followed by recruitment of factors that facilitate the repair process and finally, a re-establishment of the initial chromatin structure and eviction of the DNA damage marks from the region. Failure in this processes can result in serious predisposition to genomic damage and catastrophic consequences for the cell and the organism; due to that, our knowledge of mammalian DNA damage response would be incomplete without considering the contribution of the chromatin context and 3D organisation of the genome.

1.4.1 Access, repair, restore

As illustrated by the extensive role of chromatin remodellers in the DNA damage response, changes in chromatin conformation are essential for the repair process. There is some controversy about whether these changes are limited to the chromatin environment local to the break or whether they spread globally. Local changes have been demonstrated convincingly, using a variety of methods: HATs and HDACs are recruited to laser-induced tracks (Gong & Miller 2013) whilst high-resolution imaging of chromatin in DNA repair foci shows chromatin in a state resembling a 10 nm fibre (Dellaire et al. 2009). Consistent with this, a live cell imaging study utilising a *Sce*I/*LacO* system demonstrated local chromatin remodelling in the proximity of a break (Roukos et al. 2014). In this study, authors

used a photo-activated GFP fused to H2A, allowing them to induce damage and photo-activate chromatin within the damaged region simultaneously. They then measured changes in the H2A-GFP spot size and were able to show rapid expansion of the spot area lasting 1.5 min, followed by a re-compaction phase lasting 15 min and then hyper-condensation beyond baseline level (20-30 min). A brief local decompaction, as demonstrated in this study, would enable access of the DDR proteins to breaks. Alterations in the transcriptional activity of a locus in the vicinity of a DNA break also accompany local compaction changes. Ubiquitination of H2A at break sites was shown to correlate with transcriptional silencing near break regions (Shanbhag et al. 2010) and recruitment of the SWI/SNF remodeller PBAF is found to contribute to this silencing (Kakarougkas et al. 2014). However, a somewhat opposing finding was published in 2012, when Francia et al found evidence that transcription of small non-coding RNAs within a damaged region is required for the DNA damage response (2012). Whether the local changes in compaction and transcription spread globally is debatable. A 2006 study, using MNase digestion to assess genome-wide chromatin states, found evidence of global decondensation following DNA damage induction (Ziv et al. 2006). However a 2011 study found no evidence for global decompaction using the same approach or by sucrose gradient sedimentation to analyse the structure of soluble chromatin fibres (Hamilton et al. 2011).

Once the appropriate chromatin environment has been established, repair of the damage can proceed. The earliest step in the DDR involves rapid targeting of repair factors to the lesion and formation of DNA repair foci. The primary sensor is the MRN complex, composed of three different factors: MRE11, RAD50 and NBS1. The MRN complex activates ATM, which in turn phosphorylates H2AX at the damage site and the flanking chromatin up to a megabase away (Rogakou et al. 1999), amplifying the damage signal. An interesting question in the field is whether a full DDR is initiated only in response to DNA breaks: surprisingly, not. Tethering of early repair components to genomic regions resulted in a full DNA damage response and cell-cycle arrest, indicating that breaks are not needed beyond the initial

recruitment of factors (Soutoglou & Misteli 2008). Consistently, treatment of cells with the HDAC inhibitor TSA resulted in the activation of ATM raising the possibility that DDR can also be triggered by stimuli other than breaks, such as unusual chromatin structures (Lee 2007).

Once the necessary factors have been recruited, repair can proceed. There are two main pathways for repair of double-strand DNA breaks - non homologous end joining (NHEJ) and homologous recombination (HR). Briefly, NHEJ works by joining the ends of the break together and is active throughout the cell cycle, while in HR, which is only possible in S and G2, the non-damaged homologous locus on the sister chromatid is used as a repair template. Interestingly, some recent evidence has shown that breaks located in transcriptionally active segments of the genome are preferentially repaired with HR, while breaks in less active regions are more frequently repaired vie NHEJ even as the cells transition into S and G2 (Aymard et al. 2014). The preferential recruitment of the HR machinery to breaks in transcribed regions is found to be dependent on an interaction between the H3K36me3 mark and LEDGF, a protein that promotes HR through CtiP recruitment (Daugaard et al. 2012).

1.4.2 Transcription, replication and DNA damage

The processes of replication and transcription have been at the heart of a recent conceptual shift in the field of genome stability. While historically research on the DNA damage response was focused on external and severe mutagens such as UV light and carcinogenic drugs, recently it has become clear that DNA damage resulting from internal factors and fundamental cellular processes may be more physiologically relevant. A succession of recent studies have implicated replication and transcription as contributors to genome instability. For example a study in 2015 determined that regions of very high mutation rates within the genome overlap with Okazaki fragment junctions; the underlying mechanism was found to be retention of short segments spanning the junctions synthesised by the error-prone DNA polymerase Pol-α (Reijns et al. 2015). An earlier study identified replication

stress, physiologically present in cancer cells, as the root cause of structural and numerical chromosome instability in colorectal cancers with unstable karyotypes (Burrell et al. 2013). Transcription was implicated as a contributor to genomic instability in a publication by the Svejstrup lab-the authors found that inhibition of a transcription-associated helicase caused transcription speed to increase, resulting in recurrent chromosomal rearrangements at particular genomic regions (Saponaro et al. 2014). Another example is provided by the *RNU1*, *RNU2*, *RN5S* and *PSU1* loci, all coding for tandemly repeated, highly transcribed small RNA sequences. These four loci exhibit fragility and appear as breaks on metaphase chromosomes upon either adenovirus infection or in the absence of the Cockayne syndrome group B (CSB) protein, which is mutated in Cockayne syndrome, a rare disorder characterised by neurological and developmental defects. It has been speculated that CSB loss causes RNA polymerase stalling and blockage at the *RNU1*, *RNU2*, *RN5S* and *PSU1* loci, which then interferes with chromosome condensation and consequently, the stability of the four regions (Yu et al. 2000).

Unlike external factor-mediated instability, which usually arises from stoichiometric interactions of the damage-inducing agents with DNA and results in predictable outcomes, internally mediated instability is stochastic: it is likely to result from a combination of factors, including the exact chromatin context at the location where problems arise. While in the past most common strategies for studying the role of chromatin in genome stability involve triggering DNA damage through methods such as irradiation, laser marks or harsh damage-inducing agents such as hydroxyurea, it is clear that this new view of the field will require novel models and methods. A good model for how complex relationships between transcription, replication and chromatin influence genome stability is presented by common fragile sites (CFS).

1.5 Common fragile sites

CFS are regions of the genome prone to instability in response to replication stress, manifesting as beaks, gaps and constrictions on metaphase chromosomes. While it

is known that CFS fragility is triggered by replication stress, the exact events leading up to genomic instability are unknown. As CFS fragility is cell-type specific- different genomic locations are fragile in different cell types- it is clear that factors beyond their sequence composition contribute to fragility; in particular, replication timing and transcription are considered important, while chromatin context is a promising but under-studied potential contributor.

1.5.1 Characteristics of CFS loci

The susceptibility of cytogenetically defined genomic loci to become fragile was first identified in the 1970s on metaphase chromosomes from phenotypically normal and disease-carrying individuals (Giraud et al. 1976). Today, over a 100 fragile genomic locations known to form lesions on metaphase chromosomes have been identified in humans and classified as either rare or common fragile sites. Rare fragile sites are caused by expansion of trinucleotide repeat sequences; they are present at a low frequency in the human population (less than 5%) and show Mendelian inheritance. In contrast, common fragile sites (CFS) are found in most individuals as a part of the normal chromosomal architecture when triggered by replication stress.

Unlike rare fragile sites, CFS loci are not composed of repeat-rich sequences and features of the underlying sequences do not provide immediate clues to their fragility. However, these locations have shown a tendency for certain characteristics, which have guided theories about their fragility. The longest standing observation is that CFSs tend to be predominantly located in late replicating regions of the genome and that the timing of their replication is affected by aphidicolin. The preference of CFSs to late-replicating genomic regions was first inferred cytogenetically, via their localisation to late-replicating cytogenetic bands. An early study using FISH and immunofluorescence to define replication indicated that FRA3B is late-replicating and that its replication is further delayed in the presence of aphidicolin (Le Beau 1998). Replication changes in the presence of aphidicolin were also demonstrated for another fragile site, FRA7H (Hellman et al.

2000). Yet another FISH study, this time on the FRA6E locus, introduced a slightly different idea: the fragile region seemed to span a boundary between an early and late replicating region (Palumbo et al. 2010). The idea that CFS sites span early-late boundaries, rather than simply late regions has also been supported by a studies using consecutive staining for G-bands and R-bands to map fragile sites (Debatisse et al. 2006; El Achkar et al. 2005). With the advance of high-throughput sequencing technologies, more refined ways of mapping replication became available, including Repli-seq and small nascent strand (SNS)-seq. Large amounts of replication timing data have been generated and are now publically available. However, analysis of whether CFSs align with late replication regions on the molecular level has been prevented by the lack of cell-type matched, aphidicolin-treated data and the fact that CFSs tend to be mapped at the cytogenetic, rather than the molecular level.

Another consistent feature of CFSs across cell types is that they frequently harbour genes larger than 0.5 Mb. The tendency of CFSs to contain long genes was initially observed in lymphoblastoid cells but also held with fragile regions characterised in other cell types. More recently, a study mapping CFSs in a number of cell lines from different lineages defined a pool of possible CFS locations for all cell types, consisting of regions with genes larger than 300 kb (Le Tallec et al. 2013). An immediate question following from that observation is whether transcription of the encompassed genes matches with the cell-type specific expression of CFSs. Efforts to correlate CFS fragility with gene expression in a cell-type specific manner have given conflicting results. A 2011 study showed a correlation between expression of the FHIT gene at the FRA3B fragile site and FRA3B fragility, accompanied by an increase in R-loop formation in the presence of aphidicolin (Helmrich et al. 2011). However a more recent study from 2013 failed to find a correlation between expression and fragility on a more genome-wide scale (Le Tallec et al. 2013). The tendency for CFS regions to overlap with large genes has given rise to suggestions that their fragility may be due to interference between transcription and replication, which will be discussed in more details below.

Beyond replication timing and gene overlap, yet another similarity between CFS locations is seen at the sequence level. CFSs tend to be enriched in low complexity, A/T-rich sequences, which are associated with increased flexibility of the DNA template. A/T-rich sequences were identified as a strong predictor of fragility in a 2012 study which tried to identify features of CFS bioinformatically, selecting from multiple genomic features to build a logistic model including the features most predictive of the presence of fragile sites (Fungtammasan et al. 2012). Apart from the presence of A/T rich regions, authors identified R-bands and long distance from centromeres as other strong negative predictors, however many of the predictive genomic features are correlated and their separate influences are difficult to dissect. The enrichment of A/T sequences in fragile site regions is thought to be related to the propensity of such sequences to form secondary structures, which could cause stalling of replication forks. However, a sequence-based hypothesis of CFS fragility is bound to be incomplete, as it would not account for the cell type specific manner of CFS expression.

1.5.2 CFS in evolution

Cytogenetic studies show that the fragility of common fragile sites is conserved throughout mammalian evolution in syntenic chromosomal regions. Surprisingly, such regions of conserved fragility show a high degree of inter-species similarity at the sequence level, demonstrating that despite the local chromosomal instability, CFS do not accumulate mutations at elevated rates.

CFS have been found in corresponding locations in apes and monkeys and their locations correspond to sites of evolutionary breakpoints (Ruiz-Herrera et al. 2002). CFS regions in mice correspond to syntenic human regions and have been reported at long genes which correspond to human CFS, for example *GRID2* (associated with FRA4F), *WWOX* (FRA16D) and *FHIT* (FRA3B) (Helmrich et al. 2006). Mouse tumour cells also exhibit instability at the FHIT locus (Shiraishi et al. 2001).

At the human population level, CFS are assumed to be expressed within all individuals within a population. The main evidence for this comes from the fact that

the same cell lines derived from different individuals show fragility at similar locations. However, there is evidence that for lymphocytes at least, CFS expression levels may differ slightly between individuals, especially for locations that only show weak fragility. This was the case in a 2009 study which compared frequency of breakage at a large number of locations between three individuals (Mrasek et al. 2010). Another, earlier study, sampled nine individuals and again found differences in the frequency of breakage at CFSs, both between individuals and also between cells from the same individual sampled at a different time (Craig-Holmes et al. 1987). Overall, it appears that the most frequent CFS loci are consistently fragile between individuals, even if the breakage frequencies vary slightly.

Overall, the evolutionary conservation of CFS as a feature of chromosome make up suggests that they are not selected against, they are an inevitable feature of chromosome organisation or they are beneficial for organisms.

1.5.3 Models of CFS formation

Most ex-vivo and tissue culture cells do not form cytogenetic CFS lesions spontaneously. Instead, CFS fragility can be induced by different methods, with each method causing instability in a distinct subset of CFS. The most frequent and widely studied common fragile sites are induced by low doses of the replication inhibitor aphidicolin (APH). Used at low doses, APH does not block replication entirely, but instead causes delayed replication and replication stress. The exact molecular events that lead up to the formation of metaphase lesions at aphidicolin-induced common fragile sites are unknown.

Three models have been proposed to explain how the induction of replication stress results in genomic instability in a locus-specific manner (summarised in Figure 1-2). The “replication fork collapse” model suggests that the AT-rich sequence of CFS makes them prone to forming secondary structures which contribute to replication fork stalling and collapse (Durkin & Glover 2007). The “transcription-replication collisions” model is based on the observation that fragile sites frequently span long genes, raising the possibility that CFS instability can be the result of concomitant

transcription and replication (Helmrich et al. 2011). The "replication-initiation paucity" model explains CFS fragility as a consequence of cell-type specific features of replication timing (Letessier et al. 2011).

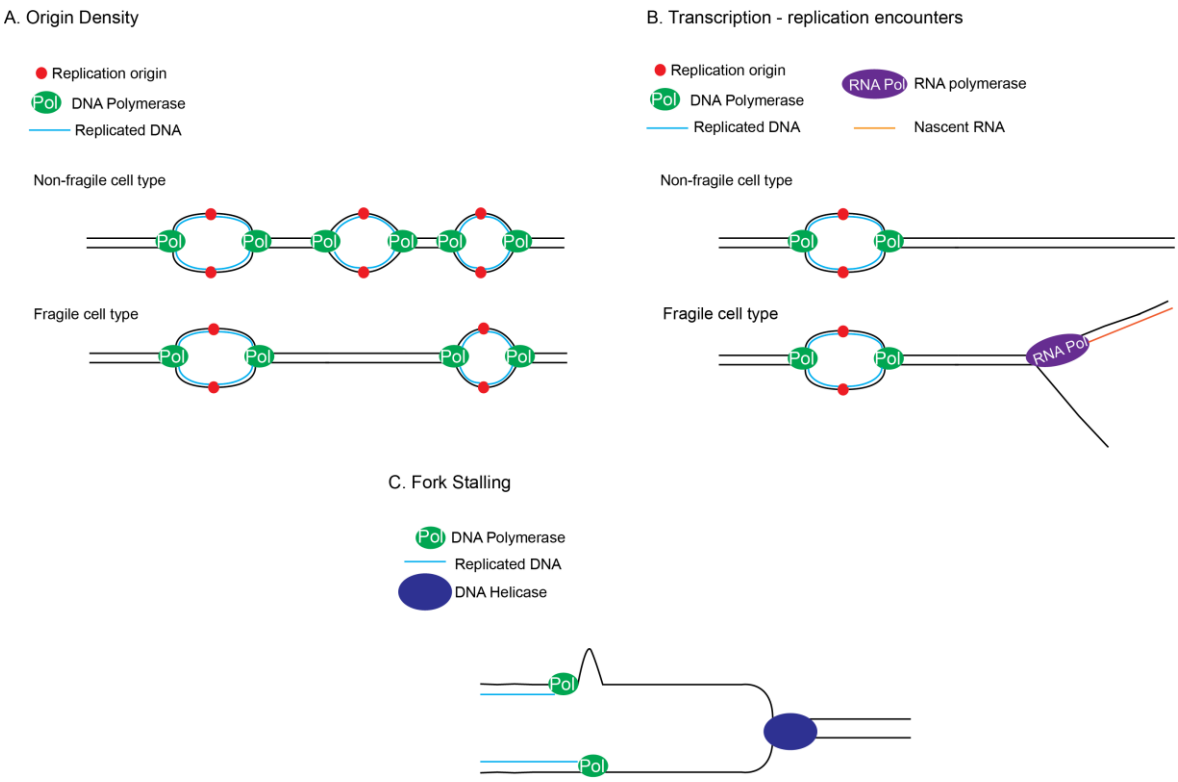


Figure 1-1 Models of CFS Formation: A. Origin density and model. Replication programmes are cell type specific. In the non-fragile cell context, the cell-type specific program specifies initiation events across the locus, avoiding instability. In the fragile context, paucity of initiation events forces replication forks to travel long distances across the locus, resulting in fragility. B. Transcription-replication collisions model. Co-occurring transcription and replication lead to CFS instability through replication perturbation and R-loop formation. C. Fork collapse model. The helicase complex travels ahead of the DNA polymerase, exposing single stranded regions that may form secondary structures. Replication fork collapses upon encountering non-B-DNA structures, leading to DNA breaks.

The fork collapse model is built on a combination of two observations. The first is that the presence of aphidicolin appears to cause uncoupling of the replicative helicase MCM2-MCM7 from the rest of the replication machinery, leaving long stretches of uncoiled, RPA-bound single stranded DNA exposed (Walter & Newport 2000). The second factor is the enrichment of flexible, A/T rich sequences at CFS,

which, when unwound and exposed by the uncoupling of the replication machinery, could form secondary structures that may result in stalling and collapse of the replication forks. The biggest flaw of this model is that it fails to explain the cell type specificity of fragile site expression and why sequences with similar A/T compositions as CFS remain stable in conditions of replication stress. Molecular data also appears to counteract the model – combing experiments at FRA6E demonstrated comparable fork speed to the rest of genome (Palumbo et al. 2010). A study of replication dynamics at the FRA3B locus also found that the speed of replication forks along the locus was not significantly different from the rest of the genome; in the presence of aphidicolin, there was a slow down of the fork and evidence of stalling at FRA3B but that was comparable to the rest of the genome (Letessier et al. 2011). In support of the fork collapse model, two genetic disorders characterised by increased fragile site formation, Bloom Syndrome and Werner Syndrome, are caused by deficiencies of RecQ helicases specialised in resolving stalled replication intermediate structures (Mohaghegh et al. 2001). Werner syndrome is caused by a deficiency of the Werner syndrome protein (WRN), an ATP-dependant helicase which efficiently unwinds structures resembling stalled replication bubbles such as Holliday Junctions (HJ). Cells derived from WRN-deficient patients form breaks at CFS spontaneously in the absence of aphidicolin treatment, while in wild-type cells, an increased frequency of CFS formation is observed following WRN depletion (Pirzio et al. 2008). BLM syndrome is caused by a deficiency of the Bloom Syndrome protein (BLM) and is characterised by increased susceptibility to early onset cancers. BLM resolves structures that mimic replication and recombination intermediates, such as HJs, via homologous repair in a manner which does not result in a crossover and BLM has been shown to localise to stalled replication forks in vivo (Sengupta et al. 2003). Cells from Bloom syndrome patients show an increased sensitivity to aphidicolin and an increased frequency of sister chromatid exchanges which could result from crossover-mediated repair of HJs by alternate nuclease complexes. Interestingly, in the absence of BLM and other Holliday junction dissolution mechanisms, extreme chromosome abnormalities

resembling multiple fragile site breaks are observed (Wechsler et al. 2011). Additional supporting evidence comes from a 2011 study demonstrating replication fork stalling at AT-rich sequences at the FRA16D fragile site, as would be expected if forks stalled frequently across the CFS sequence (Ozeri-Galai et al. 2011). Overall, evidence suggests that fork stalling is implicated in CFS breaks, but is not sufficient to fully explain the sensitivity of these sites to replication stress.

The tendency of fragile regions to encompass large genes has inspired a model suggesting that CFS instability results from collisions between the transcription and replication machinery. Large genes require longer times for transcription, sometimes exceeding the length of a full cell cycle, indicating that transcription might be ongoing during S-phase. Normally, S-phase transcription and replication are spatially separated in eukaryotic cells; most actively transcribed genes are early-replicating and changes in transcription during development are accompanied by changes in replication timing (Hiratani et al. 2009). In this model, aphidicolin treatment interferes with the temporal and spatial separation of replication and transcription at large genes, causing the occurrence of transcription and replication at fragile sites. The model speculates that concurrent transcription and replication can cause instability through the formation of RNA-DNA (R-loop) hybrids or through head-on collisions of the transcription machinery and the replication bubble, causing replication fork collapse. Efforts to correlate CFS fragility with gene expression in a cell-type specific manner have given conflicting results. A 2011 study showed a correlation between expression of the FHIT gene at the FRA3B fragile site and FRA3B fragility, accompanied by an increase in R-loop formation in the presence of aphidicolin (Helmrich et al. 2011). However a more recent study from 2013 failed to find a correlation between expression and fragility on a more genome-wide scale (Le Tallec et al. 2013). Furthermore, breaks at CFS are not restricted to transcribed regions and can also occur at intergenic sequences. Therefore unlike the RNU loci, active transcription is not required for induction of fragility at CFS suggesting that the transcription-replication collision model does not fully explain CFS lesion formation.

In the replication initiation paucity model of CFS formation, instability is caused by a cell-type specific lack of initiation events across fragile regions, forcing the forks to travel long distances to replicate CFS loci and causing the regions to remain unreplicated at the end of S phase in the presence of replication stress. Evidence supporting the model comes from a study demonstrating that a lack of initiation events across the well-studied FRA3B site correlates with its fragility in lymphoblastoid cells (Letessier et al. 2011); in contrast initiation events across the site were observed in fibroblasts, where FRA3B is stable. In addition, the authors demonstrated increased use of origins in response to aphidicolin treatment at the flanking regions, but not the core of FRA3B, showing that a failure to utilise additional origins during replication stress may also contribute to fragility. An identical analysis of replication timing across two common fragile sites in fibroblast cell lines showed a similar pattern of relatively large region, devoid of origins, where fork travel a long way to terminate over the CFS region (Le Tallec et al. 2011). Another study, using small nascent strand mapping to assay origins of replication, found four origins across a 50 kb region within the FRA3B locus but observed they were less efficient compared to origins in different parts of the genome (Palakodeti et al. 2009). Paucity of initiation events across large genomic distances is related to the so-called replication timing transition zones – areas separating domains with different replication timing. Alignment of CFS regions with such zones, based on molecular data, is consistent with a previous observation of cytogenetic CFS breaks localising at boundaries of early and late replicating chromosome bands. However, this model also appears to be incomplete, since not all genomic regions devoid of origins form active fragile sites.

To date, no model has been found to exclusively explain the cell-type specific fragility or CFS loci and it is likely that aspects of all three models contribute to CFS instability. While all models converge on the idea of replication perturbations increasing the likelihood of CSF regions remaining unreplicated or uncompacted, no full mechanistic explanations have been proposed yet. In addition, all models fail to consider a possible role for the local chromatin environment.

1.5.4 Fate of CFS in the cell cycle

Unlike other models of genomic instability, fragile site instability appears to be linked to cell cycle progression. As the cells move through the cell cycle in conditions of replication stress, fragile site regions are thought to escape the G2/M checkpoint unreplicated, giving rise to single stranded lesions. Alternatively, it is feasible that CFS regions are fully replicated in G2, but fail to compact correctly or that sister chromatids fail to separate post-replication, forming late-replication intermediates and catenanes. This idea is supported by evidence that the process of chromosome condensation and sister chromatid separation that precedes mitosis is dependent on successful and timely replication. In *S. cerevisae*, unreplicated DNA fails to condense prior to mitosis and condensin binds at replication termination sites (Dulev et al. 2008). Condensin binding at replicated sites has also been demonstrated in HeLa cells (Ono et al. 2013). The Hickson group has suggested that unreplicated CFS catenanes are marked by the DNA damage Fanconi anaemia pathway protein FANCD2 from G2 to mitosis (Chan et al. 2009). These FANCD2-marked twin foci have been observed on sister chromatids in metaphase chromosomes and can be seen to be physically linked following APH treatment. The structure-specific nuclease MUS81-EME1, which has a preference for structures resembling late recombination intermediates, co-localises with FANCD2 to sites of ongoing replication in G2/M and aids sister chromatid separation (Naim et al. 2013). Interestingly, multiple studies have reported that depletion of MUS81 results in a decrease in cytogenetically visible CFS breaks, suggesting that the structure specific nuclease is necessary for break formation. In metaphase, a mixture of breaks, gaps and constrictions can be seen at CFS loci on compacted chromosomes following replication stress. This diversity may reflect a mix of underlying structures such as double stranded DNA breaks, single stranded DNA breaks, compaction failures with uncompacted DNA bridging the gaps, and sister chromatid concatenations. In anaphase, ultra-fine bridges (UFB), DNA structures connecting daughter DNA masses, have also been linked to CFS formation. Like CFS, UFBs increase in frequency following aphidicolin treatment. A subset of UFBs are also characterised

by the presence of symmetrical FANCD2 foci at each side of the bridge, leading to suggestions that FANCD2-positive UFBs form specifically at CFS regions. This has yet to be confirmed, raising the question of whether UFBs represent a stage of CFS formation or resolution. As aphidicolin-treated cells progress through to the following G1, a subset of DNA damage response proteins form foci at increased frequency compared to non-treated cells. 53BP1, a DNA damage response protein known to promote end-joining repair of double stranded breaks, forms twin foci at fragile sites following APH treatment (Harrigan et al. 2011). This may indicate that damaged CFS foci might form breaks following mitosis and become subsequently sequestered for repair. Interestingly, under conditions of MUS81 depletion, which leads to reduction in the frequency of breaks in anaphase, the number of CFS-associated 53BP1 foci in G1 is actually increased, indicating that metaphase lesions may be a necessary step in the processing of CFS through the cell cycle (Naim et al. 2013).

As CFS appear to be processed throughout the cell cycle, it is interesting to consider how their fragility may be propagated through the cell cycle stages. Replication stress during S phase leads to mitotic problems and subsequent sequestering of CFS in 53BP1 foci in the following G1. It is easy to envisage that this damage-inducing cycle could intensify the problems at CFS loci in subsequent cell cycles under conditions of sustained replication stress.

1.5.5 CFS and disease

Although CFS are considered a normal feature of chromosomes and appear to be present in all individuals within a population, a small number of monogenic disorders are associated with increased fragility of CFSs or increased sensitivity to aphidicolin. These include the previously mentioned Bloom and Werner syndrome, as well as Seckel syndrome, caused by insufficient *ATR* levels (Casper et al. 2004).

In addition, CFSs are well known to form sites of frequent deletions translocations and amplifications in tumours. A large proportion of homozygous deletion clusters genotyped in a panel of cancer cell lines were associated with CFSs (Bignell et al.

2010). The contribution of CFS-related changes to the tumour mutational landscape is likely to be under-appreciated due to a difficulty in matching tumour types to tissue specific CFS expression patterns. Recently, an extensive mapping study found that some of the most frequent deletions in human cancers overlap with defined CFS and speculated that this overlap is likely to increase as more CFSs are mapped (Le Tallec et al. 2013). An important question is whether CFS-associated genomic changes in cancer are causal in the tumorigenesis process or just a by-product of endogenous replication stress present in cancer cells. In support of a causal relationship, some CFS overlap with known tumour suppressor genes, for example FRA3B, *FHIT*, FRA16D and *WWOX*. On the other hand, prolonged over-expression of the oncogenes cyclin E and H-Ras induced CFS breaks in BJ fibroblasts, indicating that the context of a tumour cell is sufficient to induce CFS fragility (Miron et al. 2015). Surprisingly, each oncogene induced a subtly different subset of CFS, hinting at the complex pathways behind CFS expression. In the case of over-expression of cyclin E, another study pinpointed the cause of endogenous replication stress and CFS breakage to replication in conditions of reduced nucleoside concentrations (Bester et al. 2011). Nucleoside supplementation also rescued replication-stress induced structural aberrations in CIN+ cancer cell lines, illustrating how the complex dis-regulated environment can result in instability at sensitive locations even in the absence of selective pressures (Burrell et al. 2013). Overall, the propensity of CFS to mutate in tumour cells suggests that the pharmacological induction of fragility with aphidicolin is representative of a physiological process of CFS instability.

1.6 Thesis Aims

Common fragile sites present a complex and physiologically relevant model of genomic instability. Although their appearance on metaphase chromosomes has been observed for decades, the exact reasons for their fragility and their significance have remained uncharacterised. Replication timing, transcription through long genes and difficult to replicate sequences have been the three factors implicated to date, however none of the proposed models are able to fully explain the fragility of CFS. In addition, all of the three factors have been considered

separately and never in conjunction with each other for a distinct set of fragile sites. Given the cell type specificity of fragile sites and the role of chromatin as a template for all of the processes that may give rise to instability, it is surprising that chromatin structure has never been investigated at active fragile sites. The aim of my project was to map transcription, replication timing and chromatin structure throughout the cell cycle for a set of fragile sites to determine the contributions of each of these three factors as determinants of fragile site instability.

One of the unresolved issues with regard to fragile sites is the exact nature of the lesions observed on metaphase chromosomes. While they were initially interpreted as double stranded or single stranded DNA breaks, the current pervasive idea is that they may represent regions of uncompacted DNA rather than physical breaks. A new idea is also that the cytogenetic breaks may be a necessary step in the processing of CFS, as indicated by the facts that they do not result in cell cycle arrest and that reduction in the number of lesions caused by depletion of Mus81 led to an increase in the number of CFS-associated DNA damage foci in the following G1. Therefore, I aimed to investigate the underlying chromatin structure at CFS in mitosis, both at chromosomes that carried lesions and also cytogenetically intact chromosomes. I performed these investigations by hybridising fluorescent probes mapping to fragile sites to chromosomes derived from cells exposed to replication stress and premature chromosome condensation as well as control cells.

I also characterised RNA expression levels across a number of active and inactive CFSs to determine if there is a simple linear correlation between transcription and fragility at my selected CFSs subset. Then, I used CRISPR genome editing to modify transcriptional levels of an active fragile site and determined the effects of these changes on the fragility of the locus. Although replication timing has always been seen as a strong contributor to CFS fragility, it has never been mapped in conjunction with CFS expression and in the presence of aphidicolin. I mapped replication timing in two cell types expressing characterised CFSs in the presence and absence of aphidicolin using an improved version of a previously developed

technique, Repli-seq. Finally, I investigated large scale interphase chromatin structure at active CFSs throughout the G1-S-G2 transition in the presence and absence of aphidicolin to determine if interphase chromatin structure plays a role in CFS instability.

2 Chapter 2: Materials and Methods

2.1 General Reagents, stock solutions and buffers

2.1.1 Sources of reagents

Chemicals were purchased from Sigma Aldrich, BDH Laboratory Supplies (AnalaR, VWR), Fisher Chemicals, and Amersham Biosciences (GE Healthcare). Enzymes were obtained from New England Biolabs, Promega, Roche or Life technologies. Cell culture reagents were purchased from Gibco (Invitrogen) unless otherwise stated.

2.1.2 Stock solutions and buffers

Alkaline lysis solutions:

Buffer P1: 50mM Glucose 25mM Tris pH 8, 10mM EDTA

Buffer P2: 0.2M NaOH 1% (w/v) SDS

Buffer P3: 3M KoAc pH 5.5

Buffer 1: 0.1 M Tris-HCl pH7.5, 0.15 M NaCl

Chloramphenicol: Stock solution was prepared by dissolving chloramphenicol powder in ethanol to a concentration of 25mg/ml

CuTBTA (Copper-Tris[(1-benzyl-1*H*-1,2,3-triazol-4-yl)methyl]amine) ligand complex: The CuTBTA complex was prepared by dissolving 25 mg of copper (II) sulfate pentahydrate (Sigma Aldrich, Cat No 451657) in 5 ml distilled water and mixing this solution with 116 mg of TBTA (Sigma Aldrich, Cat No 678937) dissolved in 5.5 ml DMSO.

DNA Gel Loading Buffer: 5xTBE with 40% Sucrose (w/v) and 0.25% Orange G (w/v)

FISH Hybridisation buffer: 50% deionised formamide (v/v), 10% dextran sulphate (v/v) 1% Tween 20 (v/v), in 2x SSC.

Genomic Lysis Buffer: 150mM NaCl, 0.5 % SDS (v/v) and 10mM EDTA.

Luria-Bertani (LB) Agar: Prepared by the addition of 10g of tryptone, 5g of yeast extract, 10g of NaCl and 15g of agar to 1 litre of water. It was brought to pH 7.0 by the addition of Sodium Hydroxide. Was prepared by technical services at the MRC Human Genetics Unit.

Luria-Bertani (LB) Broth: Prepared by the addition of 10g of tryptone, 5g of yeast extract and 10g of NaCl to 1 litre of water. It was brought to pH 7.0 by the addition of Sodium Hydroxide. Was prepared by technical services at the MRC Human Genetics Unit.

Phosphate Buffered Saline (PBS): Dulbecco's PBS (without Ca2+ and Mg2+) was 10mM Phosphate, 137mM NaCl and 27mM Potassium Chloride. Made from tablets purchased from Unipath (Oxford) by technical services at the MRC Human Genetics Unit.

SSC: 3M NaCl, 0.3M tri-sodium citrate, pH7.4 was prepared as a 20x stock by technical services at the MRC Human Genetics Unit.

TBS-T Buffer: 50 mM Tris, 150 mM NaCl, pH 7, 0.05% v/v Tween-20

TE: 10mM Tris HCl (pH7.6), 0.1mM EDTA prepared by technical services at the MRC Human Genetics Unit.

Triethylammonium acetate buffer (TAB): 2M trimethylamine, 2N acetic acid, pH 7.7

Tris Borate Buffer (TBE): 90mM Tris Borate, 2mM EDTA (pH 8.0) was prepared as a 20x Stock Solution by dissolving 108g of Tris Base, 27.5g of Boric Acid in 40ml of 0.5M EDTA and 960ml of water and was diluted before use. Was prepared by technical services at the MRC Human Genetics Unit.

TSE I: 20 mM Tris-HCl, pH 8.1, 2 mM EDTA, 150 mM NaCl, 1% Triton, 0.1% (v/v) SDS

TSE II: 20 mM Tris-HCl pH 8.1, 2 mM EDTA, 500 mM NaCl, 1% Triton, 0.1%(v/v) SDS

Transfer Buffer for Western Blots: 25mM Tris-Glycine (pH 8.3) with 20% methanol (v/v) 3.03g of Tris Base and 14.4g of Glycine were dissolved in 800ml of water and 200ml of Methanol was added.

2.2 Bacterial Culture

2.2.1 Media

For liquid cultures, strains were grown in Luria-Bertani (LB) -broth (Table 2-1) in a shaking incubator (250 rpm) at 37°C for ~16 h. For selection purposes, LB-broth was supplemented with ampicillin or kanamycin as required. For solid agar cultures, ampicillin-resistent cells were grown on ampicillin L-agar plates provided by IGMM Technical Services. Chloramphenicol agar plates were prepared by pouring 25 ml of L-Agar supplemented with 34 μg/ml chloramphenicol (Sigma-Aldrich, C0378-5G) in sterile petri dishes.

2.2.2 Bacterial strains

DH5α cells (Invitrogen) cells were used for routine cloning. JM109 High Efficiency Competent Cells (Promega) were used for TA-cloning. DH10B cells were supplied by BACPAC resources and used for propagation of BAC and fosmid clones.

2.2.3 Growth of BACs and fosmid clones

BAC and fosmid clones were obtained from BACPAC resources as agar stabs. A single colony streaked from a glycerol stock was inoculated into 5 ml L-broth. Bacteria were grown as described earlier but incubated at 37°C in the presence of chloramphenicol (34 μg/ml).

2.2.4 Bacterial glycerol stocks

For long-term storage of bacteria, glycerol stocks were prepared by adding glycerol to a final concentration of 20% v/v to 500 μl overnight culture and stored at -80°C.

2.2.5 Transformation of E. coli

All constructs were grown in Library Efficiency DH5α competent cells (Invitrogen). For transformation 50 μl of cells were thawed on ice. 50 ng of ligated constructs were added to the cells, followed by an incubation for 15 min on ice. The cells were heat-shocked at 42°C for 45 s and allowed to recover on ice for 2 min. 400 μl of SOC media (Invitrogen) were added to the cells and the culture incubated at 37°C for 30 min. Various dilutions of the culture were spread over L-agar plates supplemented with antibiotics as required.

2.3 DNA methods

2.3.1 Isolation of DNA from mammalian cells

For extraction of DNA, cells were resuspended in Genomic Lysis Buffer. RNase A/T1 mix (Thermo Fishcer Scientific, Cat No EN0551) was added to a concentration of 5μg/ml for RNase A and 12.5U/ml of RNase T1, followed by incubation at 37°C for 30 minutes. Proteinase K was added to 150 μg/ml and the solution was then incubated at 55°C for 2-16 h. An equal volume of Phenol: Chloroform: Iso-amyl alcohol (Invitrogen, Cat No 15593031) was added and mixed by vortex. Phases were separated by centrifugation (12,000 g, 15 min, RT) and the aqueous phase was removed to a new tube. The recovered aqueous phase was chloroform extracted with chloroform: IAA (containing chloroform: IAA in a 24:1 ratio) to remove residual phenol.

DNA was precipitated by addition of sodium acetate (pH 5.5) to a concentration of 0.3 M and 2.5 vols of ethanol, followed by 30 min to 16 h incubaton at -20°C. For small-scale extractions, 20 μg of glycogen (Sigma-Aldrich, Cat No G1508-5G) were added prior to precipitation as carrier. DNA was pelleted by centrifugation at 12,000g, 20 min, 4°C and the pellets washed with 500 μl of 70% ethanol to remove residual salt. The DNA was re-centrifuged at 12,000g, 5 min, 4°C, the supernatant

was removed and the pellet dried at RT. DNA was resuspended in TE or ultra pure water and quantified on a Nanodrop 1000 UV-VIS Spectrophotometer (Thermo Scientific) by measuring the optical density at 260 nm. Sample purity was assessed by measuring the absorbance at 260 nm and 280 nm, where a 260/280 nm ratio ranging from 1.8-2.2 was considered pure.

2.3.2 Gel electrophoresis of nucleic acids

DNA fragments were resolved by electrophoresis through a 1% UltraPure Agarose (Invitrogen, Cat No 16500-500) gel in 1 x TBE buffer, supplemented with 0.5 μg/ml of ethidium bromide (VWR, Cat No 443922U). Prior to electrophoresis, samples were prepared by adding 5 x gel loading buffer containing 5 x TBE with 40% Sucrose (w/v) and 0.25% Orange G (w/v) to a final concentration of 1x. Unless otherwise stated, 500 ng of 2-log DNA ladder (NEB, N3200S) was used as reference. Gels were visualised on a UV transilluminator.

2.3.3 Extraction of DNA from agarose gels

Fragments were resolved by gel electrophoresis as described above, but Sybr safe (Invitrogen, S33102) was used as a nucleic acid dye instead of ethidium bromide. 1 μl of 1:10 dilution of Sybr Safe was added to each sample prior to gel loading. Gels were imaged using a blue light and the gel slice containing the required DNA band was cut out with a razor blade. DNA was then extracted using QIAquick Gel Extraction kit (Qiagen, Cat#28704) according to the manufacturer instructions.

2.3.4 Purification of plasmid DNA

Small-scale isolation of plasmid DNA (typically 5 ml of overnight culture) was performed using a Qiaprep Spin Miniprep Kit (Qiagen, Cat No 27104) or E.Z.N.A Plasmid Mini Kit I (OMEGA bio-tek, Cat No D6942-01) following the manufacturer's protocol. For large-scale isolation of plasmid DNA, 200 ml of overnight culture was prepared by diluting a starter culture 1:1000 in L-Broth. Plasmid DNA was isolated using HiPure Plasmid Filter MAxiprep Kit (Invitrogen, Cat No K210016) following the manufacturer's instructions.

2.3.5 Purification of BAC and fosmid DNA

BAC and fosmid DNA was purified from 5 ml of overnight culture using a rapid alkaline lysis mini-prep as recommended by BacPac resources. Cells were pelleted by spinning at 3000 rpm, 10 min at 4°C. Cell pellets were resuspended in 600 μl buffer P1. Lysosyme was added to the resuspended cell pellets to a concentration of 5 mg/ml, followed by 5 μl of RNase A/T1 cocktail (ThermoFisher Scientific, EN0551). To lyse cells, 1.2 ml buffer P2 was added to the suspension and samples were mixed by inversion and incubated at RT for 5 min. To stop the lysis, 900 μl of buffer P3 was added and samples were incubated on ice for 5 min. Samples were centrifuged at 13,000 rpm, 4°C for 20 min. Supernatants were collected and percipitated with an equal volume of isopropanol and 0.3 M sodium acetate (pH 5.5) on dry ice for a minimum of 30 min. DNA was pelleted by centrifugation at 13,000 rpm, 20 min at 4°C. DNA pellets were washed with 70% ethanol and spun at 13,000 rpm, 15 min at 4°C. Pellets were air dried at RT and resuspended in 400 μl TE. To further clean up DNA, 400μl phenol-chloroform was added to the suspension and mixed by gentle inversion. Phases were separated by centrifugation at 13,000 rpm for 10 min at 4°C. The aqueous phases were collected and residual phenol-chloroform was removed by addition of 400 μl chloroform: IAA. Samples were spun at 13,000 rpm for 10 min at 4°C, the aqueous phase was collected and precipitated with 2.5 vol of ethanol and 0.3 M sodium acetate (pH5.5) for a minimum of 30 min on dry ice or at -20°C. DNA was pelleted by centrifugation at 13,000 rpm, at 4°C and the pellets were washed with 500 μl of 70% ethanol and spun at 13,000 rpm, 4°C for 15 min. DNA pellets were air-dried at RT and resuspended in 20 μl TE.

2.3.6 Restriction enzyme digestion

DNA was digested with restriction enzymes as needed according to manufacturer's instructions, typically in 10-20 μl reactions. Restriction digestion products were analysed by gel electrophoresis.

2.3.7 Ligation of DNA fragments

DNA ligation was performed with the Quick Ligation kit (NEB, Cat No M2200S) following the kit protocol. The vector and insert were ligated in a 1:3 molar ratio for 5 min at RT in 1x Quick Ligase Buffer (NEB) with 1 ml of the Quick Ligase enzyme in a 10 μl reaction.

2.3.8 PCR amplification of DNA sequences

Unless otherwise stated, PCR amplification was performed using Taq DNA polymerase (Invitrogen, 10342020) in a 10 to 20 μl reaction containing 10-100 ng of template DNA and 0.5 μM concentration of the required primers. Standard cycle conditions are shown in Table 2-2. Primer design was performed using Primer3 (Untergasser et al. 2012) and verified using the In-silico PCR function of the UCSC Genome Browser (Kent et al. 2002). For each primer set used, the optimal annealing temperature was established by testing a range of annealing temperatures between 55°C and 65°C.

Stage	Conditions
1)	98°C for 3 min
2)	98°C for 30 seconds
3)	61°C for 30 seconds
4)	72°C for 30 seconds
5)	Repeat Stage 2-5 for additional 29 cycles
6)	72°C for 5 min

Table 2-1 PCR program conditions

2.3.9 Real-time PCR

Real-time PCR was performed using LightCycler® 480 SYBR Green I Master Mix (Roche, Cat No 04707516001). Typically, 10 μl reactions were set up with a final primer concentration of 0.5 μM. Reactions were typically set up in duplicate or triplicate. Q-PCR was performed on a LightCycler® 480 System (Roche), using a standard 40 cycle program as per Roche master mix guidelines. Ct values were recorded and used in further analyses and melting curves were analysed for each primer set.

2.3.10 PCR purification

Depending on the expected yield and fragment size of the sample, MinElute PCR Purification kit (Qiagen, Cat No 28004) or QIAquick PCR Purifictaion kit (Qiagen, Cat No 28104) were used to purify amplification products from PCR reactions. Manufacturer protocols were followed. Samples were typically eluted in 10 μl of ultra pure water for MinElute-purified samples and 30 μl of ultra pure water for QIAquick-purified samples and stored at -20˚C.

2.3.11 Sanger sequencing of DNA

Sanger DNA sequencing was performed using standard techniques by IGMM Technical Services using appropriate primers. Dye terminator sequencing reactions were performed using BigDye (Invitrogen, Cat No 4337455) and processed on a 3730 genetic analyser (Applied Biosystems). DNA sequencing data was analysed using Sequencher 4.10.1 (Gene Codes Corp).

2.4 RNA methods

2.4.1 Purification of RNA from eukaryotic cells

Total RNA was purified from cell cultutes using the Qiagen RNeasy mini kit (Qiagen Cat# 74104) per the manufacturer's instructions. Prior to extraction, cells were typically trypsinised, washed in PBS and pelleted by centrifugation at 1300 rpm for 5 min. Cell pellets were lysed in RLT Buffer (Qiagen) supplemented with 1% (v/v) β-meracaptoethanol. 350 μl of RLT buffer were used for $5x10^6$ cells or less and 600 μl of RLT buffer were used for cell numbers over $5x10^6$. To remove DNA, on-column DNase I digestion was performed using the Qiagen RNase-Free DNase I kit. Column-bound RNA was eluted in 30 μl RNase-free water, quantified on a Nanodrop 1000 and stored at -80˚C.

2.4.2 Reverse transcription of RNA

Purified total RNA was reverse-transcribed using SuperScript II Reverse Transcriptase (Invitrogen, Cat# 18064-022). Reverse transcription reactions were typically set up with 1 μg of total RNA and 250 ng random hexamers in 20 μl reactions. cDNA was then stored at -20˚C.

2.4.3 Next Generation library preparation and sequencing of total RNA from eukaryotic cells

Ribosomal RNA was depleted from 10 μg of total RNA using the RiboMinus Transcriptome Isolation kit for human/mouse (ThermoFisher Scientific, Cat No K155002). Ribosomal depletion was verified on 1% agarose gel. Sequencing libraries were prepared from 100 ng of ribosome-depleted RNA using the NEB Next Ultra Directional RNA Library Prep Kit for Illumina (NEB, CatNo 7420S) following the manufacturer's protocol. PCR library enrichment was performed using the NEB Universal Primex and an index primer to allow multiplexing of libraries. Sequencing libraries were quantified using a Nanodrop 1000 and the fragment distribution assessed by Agilent 2100 Bioanalyzer (Agilent) using the High-Sensitivity DNA kit (Agilent). Samples were sequenced on a HiSeq 2500 system (Illumina), producing 50 bp, single-end reads.

2.4.4 RNA Sequencing analysis

Read quality was assessed with FastQC (Babraham Bioinformatics). An index of ribosomal DNA for the hg19 assembly was prepared with Bowtie 2 (Johns Hopkins University) and used to remove ribosomal reads. The remaining reads were then analysed with the Tophat-Cufflinks pipeline (Trapnell et al. 2012). First, reads were aligned to the genome using the TopHat aligner (Johns Hopkins University) with an hg19 bowtie2 index. The Cufflinks package (Johns Hopkins University) was then used to generate FPKM values for whole genes and individual transcripts. For visualisation purposes, TopHat-aligned bam files were first converted to bed files and then into bigwig files using Bedtools 2.25 (Quinlan & Hall 2010)

2.5 Protein analysis

2.5.1 Preparation of protein lysates from cell cultures

Preparation of protein lysates from mammalian cell cultures was performed using NuPage LDS sample buffer (Life Technologies, Cat No NP0007). Cells were typically grown in 6 well plates to 80% confluency, washed with PBS and lysed in 1x NuPage LDS sample buffer supplemented with dithiothreitol (DTT) to 12.5mM. 1 ml of LDS buffer was used per 10,000,000 cells. Cell lysates were then scraped into microfuge tubes, incubated at 70˚C for 10 min, sonicted and stored at -20˚C.

2.5.2 SDS-PAGE

The NuPage Novex (Life Technologies) mini gel system was used to separate whole cell protein extracts according to molecular weight. Protein extracts were thawed, briefly heated to 70˚C, spun at 15,000 rpm for 5 min and loaded on NuPage Novex 4-12% Bis-Tris gels (Life Technologies, Cat No NP0322BOX). SeeBlue Plus2 Pre-Stained Standard (Invitrogen, Cat No LC5925) was typically used as a reference. The mini-gel system tanks were filled with 1x NuPAGE MOPS SDS Running Buffer (Life Technologies Cat No NP0001) and electrophoresis was performed at 150 V for 70 min.

2.5.3 Western blotting

Whole cell protein lysates were size-separated by SDS-PAGE as described above. Following electrophoresis, size-separated proteins on the gels were transferred onto Immobilon-P polyvinylidene fluoride (PVDF) membrane (Millipore, Cat No IPVA00010), via electrophoretic transfer at 100 V for 90 min at 4°C in a transfer buffer containing 25 mM Tris base, 200 mM glycine and 20% (v/v) methanol. Prior to transfer, the PVDF membrane was hydrated by a brief wash in methanol, then rinsed in transfer buffer. Following transfer, the membrane containing the transferred proteins was washed in 1x TBS-T buffer and blocked in 5% w/v milk powder in 1 X TBS-T for 30 min at RT. The membrane was probed with primary antibodies in blocking solution for 2 h at RT or overnight at 4°C. Membranes were

then washed in TBS-T three times for 5 min and probed with secondary antibodies conjugated to horseradish peroxidase (HRP) in blocking solution for 1 h at RT. HRP bound to the membrane was detected using an enhanced chemiluminescence (ECL) detection kit (ThermoFisher Scientifict, Cat No 32106). ECL was added to membrane for 1 min and blotted to remove excess liquid, placed between two acetate sheets and exposed to photographic film (GE Healthcare Cat No 28906837). Film was developed using a Konika SRX-101A developer.

2.6 Cell Culture

2.6.1 Cell Lines

Two cell lines were primarily used for experiments unless otherwise stated. RPE-1 cells are a near-diploid retinal pigmented epithelium cell line of female origin transformed via expression of the hTERT gene. HCT116 is an epithelial colon cancer cell line derived from a male donor. HCT116 cells are chromosome instability negative (CIN-).

2.6.2 Cell growth and passage

RPE1 and HCT116 cells were cultured in Dulbecco's Modified Eagle Medium F12 (Gibco, Cat No 12500-062), supplemented with 10% foetal calf serum, 1% Pen-Strep and 1% L-glutamine. Additionally, growth media for RPE cells also contained 0.3% (w/v) Sodium Bicarbonate (Sigma, Cat. No S5761). Cells were typicaly maintained in 75 cm^2 flask at 37°C with 5% CO2 in a cell culture incubator. To passage, confluent cells were washed in PBS, trypsinised with 1x Trypsin : EDTA (Gibco, Cat No 5400054) and re-seeded, typically at 1:10 dilution. Cells were typically passaged once every 3 days.

2.6.3 Freezing cells

For freezing, an 80% confluent flask of cell culture was trypsinised, washed in PBS and pelleted by centrifugation at 1300 rpm for 5 min. The cell pellet was resuspended in 1.5 ml of freezing mix containing 90% FBS and 10% DMSO. The cell suspension was then split into three cryotubes each containing 0.5 ml of suspension and frozen at -80°C. For-long term storage, cells were transferred to liquid nitrogen.

2.6.4 Thawing cells

Frozen cell lines were rapidly thawed at 37°C and resuspended in 15 ml of growth medium in 75 cm^2 flasks. Resuspended cells were incubated at 37°C with 5% CO2 in a cell culture incubator for a minimum of 4 h or overnight. Media was then removed, the cell culture was washed in PBS and fresh media added to remove residual DMSO from flask.

2.6.5 Transfection of mammalian cell cultures

For transfection of mammalian cell cultures, cells were plated at 70,000 cells/ml in 6 well plates and grown overnight. Construct vectors were transfected using Lipofectamine 2000 (Invitrogen, Cat# 11668-019) and Opti-MEM Reduced Serum Medium (Invitrogen, Cat# 31985-070). For each transfection 1 μg of construct DNA was mixed with 400 μl Opti-Mem and 5 μl Lipofectamine-2000. To avoid aggregation of DNA and Lipofectamine-2000, the DNA was pre-mixed in 200 μl of Opti-Mem and separately, the 5 μl of Lipofectamine were mixed into 200 μl of Opti-mem. The two components were then mixed together and incubated for 20 min at RT. This transfection mixture was added to the tissue cultures in 2 ml of antibiotic-free media.

2.6.6 RNAi in mammalian cell cultures

To perform depletion of proteins using RNAi, cells were plated at 10,000 cells/ml in 6-well plates and grown overnight. RNAi oligos were transfected using RNAi Max (Invitrogen, Cat. No 13778075) and Opti-MEM Reduced Serum Medium (Invitrogen Cat# 31985-070). 0.4 nanomoles of RNAi oligos were added to 600 μl of Opti-MEM

and mixed by inversion; 16 μl of RNAiMax were also added to 600 μl of Opti-MEM and mixed by inversion. The two transfection mixtures were then mixed and incubated for 20 min at RT. 600 μl of the transfection mixture were added to each well of a 6-well plate in 3 ml of antibiotic-free growth media. The extent of depletion was assessed using qPCR and Western Blotting.

2.6.7 Synchronisation of mammalian cells

Mammalian cells were synchronised at the G1/S boundary by addition of aphidicolin to 5 μg/ml. Aphidicolin is an inhibitor of the replication machinery and is known to completely block replication at high doses (Pedrali-Noy et al. 1980). Cells were plated in 6-well plates or slide chambers at a density of 40,000 cells/ml and cultured overnight. Media containing 5 μg/ml aphidicolin was then added to the cells for 24 h to block cell cycle and retain cells at the G1/S boundary. Cells were washed in PBS and released in normal growth media. FACS analysis and immunofluorescence of cell populations at 2h, 4h, 6h, 8h and 10h following release showed that cells progressed synchronously through S-phase and into G2.

For synchronisation in G2, the CDK1 inhibitor RO-3306 was used. Cells were plated into 6-well plates and slide chambers at a density of 40,000 cells/ml and grown overnight. Media containing RO-3306 at 9 μM was added for 20 h to arrest cells in G2. Cells were then released in fresh media and FACS analysis showed that they progressed synchronously through mitosis and G1 for the next 10 h.

2.6.8 Immunofluorescence

Immunoflourescence was routinely performed to assess the levels of expression and exact localisation of proteins of interest. Cells were seeded on Superfrost + slides (Thermo Scientific Cat No J1800AMNZ) placed in slide chambers containing 5 ml of media at a suitable density. For fixation, slides were washed with PBS and treated with 4% paraformaldehyde (PFA, Sigma Aldrich ,Cat 158127) for 10 min at RT. Slides were washed in PBS for 5 min at RT and permeabilised in 0.02% Triton X-100 (Sigma Aldrich, Cat No T8787) in PBS for 10 min at RT. Following

permeabilisation, slides were stored in PBS at 4°C or blocked and stained immediately. Blocking was performed by incubating the slides in PBS containing 5% donkey serum for 15 min at RT in a dark humidified chamber. Following blocking, primary antibodies were added onto the slides at the required dilutions in blocking solution and incubated at RT for 1 h in a dark, humidified chamber. Slides were then washed in PBS / 0.01% Tween three times for 5 min at RT. Secondary antibodies, raised in donkey and conjugated to fluorophores (Jackson Immuno Research), were diluted 1:500 in a blocking solution, added to the slides and incubated for 1 h at RT in a humidified chamber. Slides were washed in PBS containing 0.01 % Tween-20 three times for 5 min at RT. To detect DNA and nuclei, slides were stained in 50 μg / ml DAPI for 3 min at RT. Slides were then drained and mounted in Vectashield (Vector Laboratories, Cat No H-1000). Imaging was performed on a Zeiss Epifluorescence microscope using 100x objective.

2.6.9 EdU staining of mammalian cells

Click-it flow cytometry kit (Invitrogen, Cat No C10634) was used to visualise active replication in mammalian cells following the manufacturesrs instructions. To visualise sites of active replication, cells grown on slides were pulsed with the thymidine analogue 5-ethynyl-2'-deoxyuridine (EdU). EdU was added to exponentially growing cell cultures at a concentration of 5 μM for 30 min in HCT116 cells and 1 h in RPE1 cells to account for differences in the cell cycle dynamics between the two cell types. Slides were washed in PBS and fixed in 4% PFA for 10 min. To remove residual PFA, slides were washed in PBS for 3 x 5 min, at RT. Permeabilisation was performed in PBS / 0.2% Triton for 10 min, after which slides were washed again in PBS 3 x 5 min at RT. Slides were then incubated at RT in a dark humidified chamber in the presence of a click reaction mixture, prepared as per manufacturer's instructions, which included copper sulphate and fluorescently labelled azide, enabling the cyclo-addition of the fluorescently labelled azide group onto the alkyne group of EdU. Following a thirty-min incubation with the click reaction mixture, the slides were washed in PBS 3 x 5 min to remove un-clicked fluorescent azide groups. Following the click labelling, slides were typically stained

in a solution containing 50 μg / ml DAPI for 3 min at RT to enable visualisation of cell nuclei. Slides were mounted in Vectashield and imaged on a Zeiss epifluorescence microscope using 100x objective. Replicating cells, positive for EdU, displayed replication-stage dependant EdU localisation patterns corresponding to patterns previously described in the literature for another thymidine analogue, bromodeoxyuridine (BrdU) and components of the active replication machinery such as PCNA (Dimitrova & Gilbert 1999) . Early replicating cells showed a diffuse pattern of staining, while in late - S EdU signal showed a focal distribution with high levels of signal at the nuclear periphery and heterochromatic regions.

2.7 Flow cytometry analysis and sorting of mammalian cells

2.7.1 Sorting of cells expressing GFP

To sort cells expressing GFP, cell cultures were trypsinised and resuspended in growth media at a density of $1x10^6$ cells/ml. For analysis, an Accuri C6 analyser (BD Biosciences) was used and GFP-generated fluorescence was measured in the FL1 channel. A non-GFP expressing cell population was used to determine a negative population gate prior to analysis. For single-cell soring, the cell population was run through a FACSJazz sorter (BD Biosciences). Fluorescense was measured in the FITC/GFP channel and a non-transfected, GFP-negatve population was used for gating.

2.7.2 Cell cycle assessment and sorting using propidium iodide staining

Propidium iodide (PI) is a fluorescent, intercalating nucleic acid dye which shows a 20-30 times increase in fluorescence when bound to nucleic acids. PI staining was used to analyse the cell cycle distribution of a cell population and sort cells based on cell cycle stage. The cell population to be analysed was trypsinised, pelleted by

centrifugation at 1300 rpm for 5 min and washed in PBS. Cell pellets were then resuspended in PBS at a density of 1.5 x 10^6 cells/ml and ethanol was slowly added to the cell suspension to a concentration of 70% to fix and permeabilise the cells. Cells were then incubated on ice for a minimum of 30 min or stored at 4°C for up to 2 weeks. For PI staining and cell cycle analysis, ethanol-permeabilised cells were pelleted by centrifugation at 1800 rpm for 5 min at RT and washed in PBS. The cell pellets were then stained in a solution containing 1 μg/ml PI and 4μg/ml RNase A in PBS at 2 x 10^6 cells/ml for a minimum of 30 min at RT.

Cell cycle analysis was performed on a LSR Fortessa analyser (BD Biosciences) by measuring the fluorescence in the 695/740 channel. For PI based-cell sorting, nuclei were isolated from the fixed, permeabilised cells via pepsin digestion, as described in ((Dileep et al. 2012). Briefly, cells were pelleted, washed in PBS and resuspended in 0.015% pepsin (Sigma, CatNo P6887) and 0.01N HCl at a concentration of 400,000 cells/ml. Pepsin digestion was allowed to proceed for 20 min at 37°C before cells were pelleted by spinning at 600 g for 5 min at RT. Cells were stained in a solution containing 1μg/ml PI and 4 μg/ml RNAse A in PBS at a concentration of 2x10^6 cells/ml. Staining was performed at RT, for a minimum of 30 min and maximum of 2 h. Sorting was typically performed on a FACSAria sorter (BD Biosciences) by measuring the fluorescence in the 685/735 channel and gating the cell population as required. Subsequent analysis and visualisation of FACS-analysed or sorted samples was performed using Flow-Jo software.

2.7.3 PI/EdU analysis of cell cycle

Dual PI and EdU staining was performed to simultaneously determine the cell cycle stage and the proportion of actively replicating cells within a population. EdU staining was performed using the Click-iT Plus EdU Alexa Fluor 647 Flow Cytometry Assay Kit (Invitrogen, CatNo C10634) following the manufacturer's instructions. Exponentially growing cell cultures were pulsed with 5 μM EdU for 30 min or 1 h (HCT116 and RPE1 cells, respectively). Cells were trypsinised and resuspended in 300 μl PBS per 5x10^6 cells. Cells were fixed by the slow addition of 700 μl ethanol

and incubated on ice for a minimum of 30 min or stored at 4°C for up to two weeks. For staining, cells were washed in PBS and permeabilised in 1 x saponin permeabilisation reagent (Invitrogen) for 20 min at RT. Next, cells were resuspended in a click reaction mixture following the manufacturer's instructions. The reaction mixture contained copper protectant (Invitrogen) and an azide conjugated to an Alexa 647 fluorophore (Invitrogen), which allowed the addition of the fluorescent group on the EdU molecule and its subsequent visualisation. 1 ml of click reaction mixture was used per 1 x 10^7 cells and the click reaction was allowed to continue for 30 min at RT, protected from light. Following the click reaction, cells were washed in 0.5 ml 1 x Saponin reagent and stained in a mixture containing 1 μg/ml PI and 4 μg/ml RNAse A at a concentration of 2 x 10^6 cells/ml for 30 min at RT. Cells were typically analysed on a Fortessa analyser by measuring the signals in the 695/740 channel for PI and the 730/745 channel for Edu/Alexa 647.

2.8 Fluorescent in-situ hybridisation (FISH)

2.8.1 Preparation of human metaphase chromosomes

Human metaphase chromosomes were prepared from cultures of exponentially growing cells. To induce mitotic arrest and increase the numbers of mitotc cells, colcemid, (Life Technologies, Cat No 15210-040) an inhibitor of microtubule depolymerisation, was added to the cell cultures to a concentration of 0.1 μg/ml. RPE1 cells were treated with colcemid for 1 h prior to harvest, while HCT116 were only arrested for 30 min to account for differences in the speed of cell cycle progression between the two cell types. Following colcemid treatment, cells were harvested, washed in PBS and resuspended in 5ml hypotonic solution, containing 75 mM KCl. The hypotonic treatment of cells was performed at RT for 10 min, after which cells were pelleted by centrifugation at 1200 rpm for 5 min. Next, cells were fixed in 5 ml of freshly prepared solution of 3:1 ratio methanol: acetic acid (MAA). The MAA fixative was added to the cell pellet dropwise with constant agitation. Cells were pelleted by centrifugation at 1500 rpm for 5 min and the MAA fixation

step was repeated two more times. Chromosome preperations were stored at -20°C.

To prepare slides with metaphase spreads, metaphase chromosome preparations were dropped onto glass slides. The glass slides were pre-treated in a dilute solution of HCl in ethanol for at least an hour prior to use. The chromosome preparations were pelleted by centrifugation at 1500 rpm for 5 min and resuspended in freshly prepared MAA solution until the suspension became cloudy. Two drops of the suspension were dropped onto a pre-treated glass slide from a height of 20c m and dried at RT overnight before staining or hybridisation.

2.8.2 Preparation of FISH probes

BAC and cosmid FISH probes were isolated as described in 2.3.5. Prior to use in a hybridisation, probes were labelled using a nick translation reaction with the uridine analogues biotin-16-dUTP (Roche, CatNo 11093070910) or digoxigenin-11-dUTP (Roche, CatNo 11093088910). Nick translation was performed in a 20 μl reaction volume, containing 1-1.5 μg DNA with 5 μl each of 0.5 mM dATP, dCTP and dGTP and either 2.5 μl of 1 mM biotin-16-dUTP or 1 μl of 1 mM digoxigenin-11-dUTP. DNase I (Roche, Cat No 4716728001) was added to a final concentration of 1 U/ml and DNA polymerase I (Invitrogen, Cat No 18010025) was added to a final concentration 0.5 U/μl. The reaction was performed in 1 x nick translation salts (NTS) buffer, containing 50 mM Tris pH7.5, 10 mM $MgSO_4$, 0.1 mM DTT and 50 μg/ml BSA for 90 min at 16°C. The reaction volume was then increased to 80 μl by addition of 60 μl TE and unincorporated nucleotides were removed by gel filtration of the NTS reaction through a G50 Sephadex spin column (Roche, Cat No G50DNA-RO) following the manufacturer instuctions. Labelled probes were then stored at RT, protected from light.

2.8.3 Quantification of Label Incorporation

The amount of incorporated labelled nucleotides in each probe was quantified by spotting the labelled probes on a nitrocellulose membrane and probing the

membranes with alkaline-phosphatase conjugated streptavidin or anti-digoxigenin antibodies. Nitrocellulose membranes (Protran, Whatman, Cat No BA 85120) were soaked in water for 5 min, followed by immersion in 20 x SSC for 10 min. Membranes were allowed to dry and stored at RT until use. DNA from labelled probes was spotted onto the membranes in dilutions of 1/500, 1/1000, 1/5000 and 1/10 000 in TE. Brightly labelled probes were also spotted on the membranes in similar dilutions to serve as controls. DNA was crosslinked to the membrane by 150 mJ of UV irradiation. Following crosslinking, the membrane was briefly washed with Buffer 1 and blocked in 3% BSA w/v in Buffer 1 for 30 min at 60°C. The membrane was probed with either strepavidin-alkaline phosphatase, and/or anti-digoxigenin-alkaline phosphatase Fab fragments (Roche, Cat No 11093274910 and 11093266910), diluted 1:1000 in Buffer 1 for 2 h at RT. Excess antibody was removed by washing 2 x 15 min at RT with Buffer 1, after which time the membrane was equilibrated for 5 min in 0.1 M Tris-HCl (pH 9.5). The membrane was developed in a small sealed plastic bag containing two drops from components 1-3 of the alkaline phosphatase substrate kit IV (Vector Laboratories, Cat No SK-5300). The substrates in this colour reaction are 5-bromo-4-chloro-3-idolyl phosphate and nitroblue tetrazolium, producing a blue reaction product in the presence of alkaline phosphatase. A complete colour reaction could be observed within a few h to assess whether the probes were suitably labelled.

2.8.4 Hybridisation of FISH Probes

To prepare for hybridisation, slides were treated with 100 µg/ml RNaseA (Invitrogen, Cat No 12091039) in 2 x SSC for 1 h at 37°C, washed briefly in 2 x SSC and dehydrated through an ethanol series (2 min each in 70%, 90% and 100% ethanol). Slides were then air dried and baked at 70°C for five min before denaturing. Denaturation was performed in 70% formamide (v/v) in 2 x SSC (pH 7.5). Slides containing MAA-fixed chromosome spreads were denatured at 70°C for 1 min, while slides on which cells were cultured and then fixed in 4% PFA were denatured at 80°C for 20 min. Following denaturation, slides were submerged in

ice-cold 70% ethanol for 2 min and then dehydrated through 90% and 100% ethanol for 2 min each at RT.

To prepare probes for hybridisation, 150 ng of labelled probe were combined with 5 μg of salmon sperm and 10 μg of human Cot1 DNA (Invitrogen, Cat No 15279011). Two volumes of ethanol were added and the probe mix was spun down and dried under a vacuum. The dried probes were resuspended in 10 μl of FISH hybridisation buffer containing 50% formamide (v/v), 1% Tween-20 and 10% dextran sulphate (Sigma Aldrich, Cat No D8906-100G) in 2 x SSC. The dried probes were dissolved in hybridisation buffer for a minimum of 2.5 h at RT wiith occasional agitation. Probes were then denatured at 70°C for 5 min and reannealed at 37°C for 15 min and chilled on ice. The probes were pipetted on 22 mm square cover slips and picked up with the pre-treated slides. Commercially available chromosome paints (Cytocell), supplied in own hybridisation buffer in a ready-to-use format were also used. They were directly pipetted on the coverslips and picked up with slides. Liquid rubber (Tiptop) was applied around the coverslips and hybridisation performed in an enamel tray at 37°C in a waterbath overnight.

2.8.5 Washing and detection of FISH signal

After overnight hybridisation, the rubber seal was peeled off from the slides and the coverslips. The coverslips were allowed to float off in 2 x SSC and slides were then washed four times in 2 x SSC at 45°C for 3 min. Another four washes were then performed, with 0.2 x SSC at 60°C for 3 min. Slides were briefly transferred to 4 x SSC with 0.1% Tween 20 (v/v) and blocked in 5% milk in 4 x SSC for 5 min at RT. Antibodies were diluted as required in 5% milk powder (w/v) in 4 x SSC and then centrifugated at 13,000 rpm for 10 min to remove clumps. Detection of biotin label was performed with sequential layers of flourescin (FITC)-conjugated avidin, biotinylated anti-avidin and a further layer of FITC-avidin. DIgoxigenin was detected with sequential layers of Rhodamine-conjugated anti-digoxigenin and Texas-Red (TR) –conjugated anti-sheep IgG. The antibodies used, relevant dilutions and suppliers are listed in Table 2-3. Slides were incubated with each antibody layer for

30 min in a moist chamber at 37°C. Between antibody layer incubations slides were washed 4 times in 4 x SSC 0.1% Tween 20 for 3 min at 37°C. Following the last antibody layer, slides were washed 4 times in 4 x SSC 0.1% Tween 20 for 3 min at 37°C and then stained in a solution containing 50 μg/ml DAPI in PBS for 3 min at RT. Slides were mounted in Vectashield and coverslips sealed with rubber solution (Pang) or nail varnish. Slides were imaged on a Zeiss epifluorescence microscope using a 100x objective.

Antibody	Dilution	Supplier
Fluorescine avidin	1/500	Vector Labs (A-2011)
Biotin anti-avidin	1/100	Vector Labs (BA-0300)
Texas Red anti-sheep	1/100	Vector Labs (TI-6000)
Anti-dig rhodamine	1/20	Roche (112077509101120)

Table 2-2 Antibodies used for detection of FISH signal

2.8.6 Genomic Clones Used for FISH

A variety of genomic clones were used as FISH probes throughout the project and they are listed, along with the corresponding genomic locations in Table 2-4. For all probes, the correct cytogenetic localisation of the probe was verified by hybridisation to metaphase chromosomes prior to use in interphase cells.

Probe	Genomic Location: Start	Genomic Location: End	Genomic Band	Type
RP11-436I1	136818800	137001238	2q22.1	BAC
RP11-236P10	141182186	141337859	2q22.1	BAC
RP11-15G12	142948672	143108539	2q22.2	BAC
RP11-56K5	145006565	145166970	2q22.3	BAC
RP11-952E8	70540713	70710146	1p31.1	BAC
RP11-644A16	71956202	72140823	1p31.1	BAC
RP11-357C16	68915438	69105214	1p31.1	BAC
RP11-452B11	69176951	69356842	1p31.1	BAC
RP11-482A14	69399135	69576878	1p31.1	BAC
RP11-44E15	69595766	69781569	1p31.1	BAC
RP11-795A13	69851036	70025173	1p31.1	BAC
RP11-1085J6	139594407	139768017	4q31.1	BAC
RP11-1066F2	139856514	140068112	4q31.1	BAC
RP11-121K15	140158813	140344775	4q31.1	BAC
RP11-102K5	140710991	140879524	4q31.1	BAC
RP11-876B4	141116214	141291804	4q31.1	BAC
RP11 -667C5	141384091	141554894	4q31.1	BAC
RP11-104N8	141775105	141937724	4q31.1	BAC
RP11-57O8	142285341	142467294	4q31.1	BAC
RP11-640M2	163755124	163910201	4q32.2	BAC
RP11-946L12	164101959	164282405	4q32.2	BAC
RP11-780E8	164450766	164601741	4q32.2	BAC
RP11-47H6	165289180	165432768	4q32.2	BAC
RP11-153D1	165720700	165883800	4q32.2	BAC
RP11-776O4	113403650	113576915	3q13.2 - 3q13.31	BAC
RP11-52N10	114479903	114636091	3q13.31	BAC
RP11-11F11	115446092	115606348	3q13.31	BAC
RP11-354H5	116410230	116551551	3q13.31	BAC
RP11-696L1	117376433	117552791	3q13.32	BAC
RP11-456O4	118915065	119071586	3q13.32 - 3q13.33	BAC
RP11-120C5	87485968	87659888	4q21.3	BAC
RP11-1053C2	89213170	89389643	4q22.1	BAC
RP11-44A17	91534737	91688609	4q22.1	BAC
RP11-351L22	92912674	93073948	4q22.1	BAC
RP11-479E18	94962968	95121208	4q22.2 - 4q22.3	BAC
RP11-155A18	76612727	76795175	7q11.23	BAC
RP11-973N8	77475501	77665434	7q11.23 - 7q21.11	BAC
RP11-7N3	79600618	79780358	7q21.11	BAC
RP11-614C5	81780834	81945425	7q21.11	BAC

RP11-649L8	83563011	83722404	7q21.11	BAC
RP11-51C21	86008317	86165627	7q21.11	BAC
RP11-975J8	87523950	87713836	7q21.12	BAC
RP11-624N7	68576767	68749059	1p31.3	BAC
RP11-915N9	86463555	86626921	4q21.23	BAC
RP11-688G4	96196712	96376784	4q22.3	BAC
RP11-6L24	97091291	97250657	4q22.3	BAC
G248P8923F10	69453824	69492900	1p31.2	Fosmid
G248P86197B3	68588964	68629214	1p31.3	Fosmid
G248P85943H10	68367192	68406422	1p31.3	Fosmid
G248P83504C1	69801869	69843403	1p31.1	Fosmid
G248P85730E10	138827973	138869427	4q28.1	Fosmid
G248P8075B1	139935150	139978608	4q31.1	Fosmid
G248P8117C5	141003415	141049416	4q31.1	Fosmid
G248P8526E8	142031870	142076104	4q31.2	Fosmid
G248P8183F5	60950550	60993172	3p14.2	Fosmid
G248P89337E4	59423151	59459364	3p14.2	Fosmid
G248P8923G12	57947227	57989923	3p14.3	Fosmid
G248P89726G8	62413310	62450319	3p14.2	Fosmid
G248P8027H7	78087080	78125909	16q23.1	Fosmid
G248P8890B4	76641081	76681724	16q23.1	Fosmid
G248P81697G2	75142199	75183028	16q23.1	Fosmid
G248P87155E4	73670724	73714216	16q23.1	Fosmid
G248P800778H5	91076048	91117811	4q22.1	Fosmid
G248P85108G3	92058557	92098840	4q22.1	Fosmid
G248P84123D7	93077304	93115635	4q22.1	Fosmid
G248P8189B8	94069299	94108069	4q22.2	Fosmid
G248P86724A11	162609737	162647264	4q32.2	Fosmid
G248P88037D8	163640717	163680360	4q32.2	Fosmid
G248P85162A12	164566239	164608074	4q32.3	Fosmid
G248P86865C5	165481486	165522185	4q32.3	Fosmid

Table 2-3 BAC and fosmid FISH probes

2.8.7 Investigation of large-scale chromatin compaction using FISH

The use of differentially labelled FISH probes to investigate large-scale chromatin structure was proposed and pioneered in 1992 (van den Engh et al. 1992) and has been used in multiple studies. Generally, two differentially labelled fosmid probes, spaced 50 kb to 2 Mb apart are hybridised and the physical distance between them measured in a large number of nuclei. The distances between probes separated by

less than 50 kb or more than 2 Mb of genomic distance are thought to be uninformative (van den Engh et al. 1992). Fosmid probe pairs separated by 0.8 Mb- to 1.5 Mb were used in this study. The correct genomic location for each fosmid was verified by hybridisation to metaphase spreads as described in 2.7.4. The distance between fosmid probes was measured in 50 to 100 nuclei. Three-channel images were taken for each nucleus, including a DAPI (blue) channel defining the boundaries of the nucleus and a FITC (green) and a TxRed (red) channel, defining the locations of the fosmid probes. To calculate the distance between the two probes, a script developed by P. Perry (Chambeyron & Bickmore 2004) was used. The area of the nucleus was calculated by segmenting the DAPI signal. After user identification of the probe pairs and automatic background reduction, segmentation was performed for the green and the red channels and the coordinates of the centroids of the probe signals were determined. Distance between the two centroids was calculated by trigonometric equations. The distances were calculated in pixels and converted to microns (multiplication factor of 0.134). Distances were normalised to account for differences in the nuclear size of the cells they were derived from. The script outputed a nuclear area value following the segmentation procedure and the nuclear radius was calculated by calculating the square root of the nuclear area value over π. The mean radius of the whole population was then calculated and a normalised radius was calculated for each cell by dividing the radius over the mean radius. A normalised distance was then derived by calculating the distance in microns/normalised radius. Unless otherwise stated, normalised distances were used in further statistical analysis.

2.9 Mapping replication timing using Click-seq

Repli-seq is a technique developed to investigate the replication timing at particular genomic regions or throughout the genome. The technique was first developed in 2010 (Hansen et al. 2010) and was originally based on labelling newly replicated DNA with the thymidine analogue BrdU. In the classic Repli-Seq technique, cycling

cells are pulsed with BrdU for a short length of time, and then sorted through flow cytometry into different S-phase fractions. DNA is then extracted from each of the different S-phase populations and newly synthesised DNA is enriched via an anti-BrdU antibody pull-down. The newly synthesised DNA from each fraction can then be hybridised on arrays, sequenced or specific regions can be interrogated using qPCR. I developed a new version of the Repli-seq approach, utilising the thymidine analogue EdU instead of BrdU and developing a new method for sequencing for the enriched DNA. The main advantage of using EdU compared to BrdU is because the EdU molecule can be stably and specifically attached to a biotinylated azide molecule using click chemistry. Biotinylated newly replicated DNA can then be enriched by streptavidin pull-down, a technique which is more robust and specific compared to the antibody based pull-down employed in BrdU-based Repli-seq. In addition, the streptavidin pull-down avoids the use of harsh denaturing treatment of DNA which is needed in the antibody pull-down to reveal the BrdU epitope. I have called this new methodology Click-seq and its details are discussed below, while the optimisation of the method is discussed in Chapter 4.

2.9.1 DNA preparation for Click-seq

Actively cycling cell cultures were pulsed with 10 μm EdU for 30 min (HCT116 cells) or 1 h (RPE1 cells). Cells were then trypsinised, washed in PBS and fixed in 70% ethanol as described in 2.6.2. Following fixation, nuclei were prepared from the cells and PI stained following the procedure outlined in 2.6.2. Nuclei were then sorted on a FACSAria flow cytometer (BD Biosciences) into three fractions: early, mid and late - S phase populations. To keep sorting conditions similar between samples with differing levels of PI staining, the sorting gates were determined by separating the area between the middle of the G1 peak and the middle of the G2 peak into three equal-sized windows, corresponding to early -, mid - and late -S phase populations. Cells were sorted into PBS supplemented with 0.25% BSA and a minimum of 150,000 cells were sorted for each population. Following the sort, nuclei were pelleted by centrifugation at 2000 rpm for 20 min. The pellets were resuspended in 400 μl of Genomic Lysis Buffer, RNase A/T1 cocktail was added and

samples were incubated at 37°C for 30 min. Proteinase K was added to 150 μg/ml and the samples were incubated at 55°C for 2-16 h. Genomic DNA was then fragmented by sonication. Fragmentation of samples intended for use in qPCR reaction was performed in a Bioruptor (DIagenode) for 18 min, 30 s on and 30 s off, with the samples suspended in a 4°C waterbath for the duration. Fragmentation of samples for next generation sequencing was performed on a Soniprep 150 probe sonicator, for 18 min, 30 s on and 30 s off at an ampliutude of 6μm. To avoid over-heating, the sample was submerged in ice for the duration of the sonication. The fragmentation of the samples was checked on an agarose gel. While the Bioruptor-fragmented samples showed a lot of variability and a wide rage of fragments, the probe sonicator showed a reproducible fragmentation pattern, with all samples showing a tight distribution between 100 and 300 bp (Figure 2-1). DNA was purified from the sonicated fractions by phenol-cloroform extraction (Section 2.3.1).

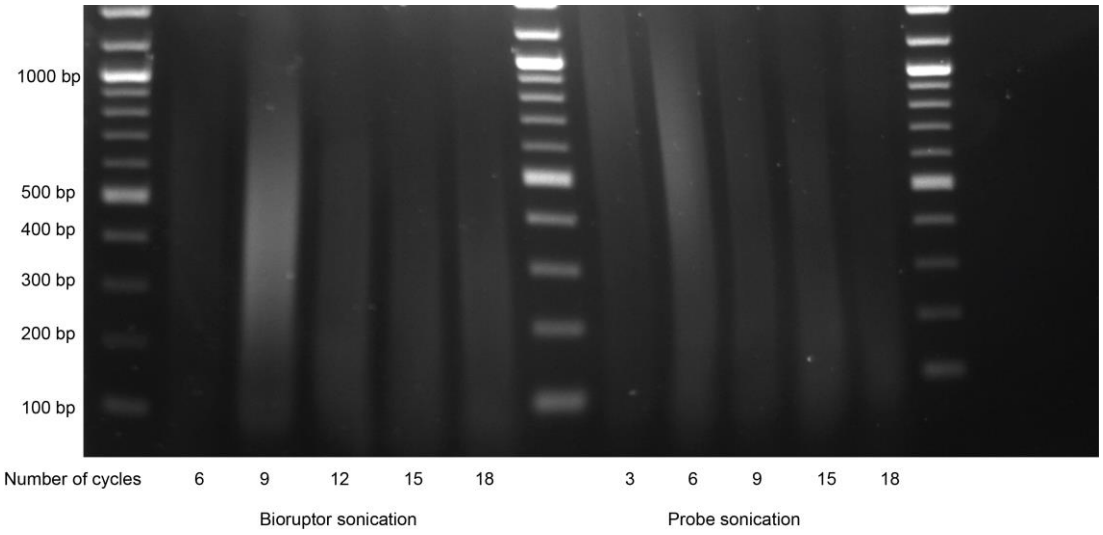


Figure 2-1 Comparison between Bioruptor and probe sonicator. Genomic DNA was sonicated for different number of cycles using the two methods. The sonicated fractions were then run out on a gel. 18 cycles on the probe sonicator were used to produce samples for the Click-seq methodology.

2.9.2 Addition of biotinylated azide using click chemistry

Click-chemistry was employed to attach a biotinylated azide to the alkyne groups of EdU molecules, labelling all newly synthesised DNA with a biotin tag. The click reaction conditions were based on a protocol recommended by Lumiprobe. The

reaction was performed with up to 3 μg of DNA resuspended in 18 μl of water in a 60 μl reaction volume. The remaining components of the click reaction were added sequentially, always following the order outlined below and in Table 2-5.

1. 3 μl of 2 M triethylammonium acetate buffer, pH 7.0
2. 30 μl DMSO
3. 6 μl of freshly prepared 5 mM ascorbic acid, which was added to the reaction to neutralise free radicals.
4. 1.2 μl of 10mM biotin azide (Life Sciences, CatNo B10184)
5. 3 μl of 10 μM CuTBTA ligand complex. The copper in the mixture catalyses the click reaction, while the TBTA ligand prevents DNA damage resulting from free copper in the reaction.

The reaction was mixed by vortexing and incubated at RT overnight, protected from light by aluminium foil. Following overnight incubation, the biotinylated DNA was ethanol-precipitated with 2.5 volumes of ethanol in 0.3 M sodium acetate with 20 μg glycogen as a carrier. Precipitation was usually performed on dry ice for a minimum of 30 min. DNA was pelleted by centrifugation at 13,000 rpm for 20 min at 4°C, washed with 500 μl of 70% ethanol and centrifuged again at 13,000 rpm for 15 min at 4°C. The DNA pellet was dried at RT for a few minutes and resuspended in 20 μl of TE. 3 μl of 'clicked' DNA was used for quantification and verification of biotin attachment (described in 2.8.4), whilst the remaining 17 μl were used in subsequent experiments.

Component	Order of addition to reaction	Volume	Concnetration in reaction	Stock details
DNA		18 μl	Up to 3 mg	
TAB Buffer	1	3 μl	100 μM	2 M stock prepared by mixing 2.78 ml triethylamine, 1.14 ml acetic acid and 6.08 ml ddH2O. Stored at RT indefinitely
DMSO	2	30 μl	50%	
Ascorbic acid	3	6 μl	0.5 mM	Freshly prepared 5 M stock
Biotin azide	4	1.2μl	0.2 mM	10 mM in DMSO
Cu/TBTA	5	2.5μl	0.5 mM	10 mM stock prepared by preparing 10 mM solution of copper II persulphate in ddH2O and 10 mM TBTA solution in DMSO and mixing equal volumes of the two solutions

Table 2-4 Click reaction components

2.9.3 Enrichment of biotinylated DNA by streptavidin pull-down

Biotinylated, newly replicated DNA was purified from total DNA by enrichment with streptavidin magnetic beads. 50 μl of C1 streptavidin Dynabeads (Invitrogen, Cat No 65001) were used per sample. Beads were pre-washed three times in 1 ml TE for 5 min at RT with rotation. The volume of the clicked DNA sample was boosted to 1 ml and salt concentration adjusted to 0.15 M NaCl. For samples to be used in qPCR reactions, 100 μl was saved as input. The sample was then mixed with 50 μl of pre-washed beads and 1 μl of Triton X-100 (Sigma Aldrich, Cat No X100) was added to prevent binding of the beads to the walls of the plastic microfuge tubes. The sample-bead mixture was incubated overnight at 4°C with rotation. Beads, bound to biotinylated DNA were separated using a magnetic rack and the supernatant, containing non-biotinylated DNA, was discarded. The beads were then sequentially washed twice with 1 ml each of TSE I, TSE II and TE. For each wash, beads were transferred to a clean microfuge tube and washed for 3 min at RT with rotation. Following the final TE wash, biotinylated DNA was eluted in 50 μl of elution buffer at 98°C for 10 min with occasional agitation. Two elution buffers were tested: water

and 95% formamide. While manufacturer instructions for Dynabeads recommend elution in 95% formamide, water was also tested to determine if degradation of DNA associated with boiling in formamide can be avoided. Side by side elutions of identical samples found that similar amounts of material are eluted in water and formamide. Therefore, water was predominantly used to elute biotinylated DNA, although some early experiments were performed with formamide elutions. When formamide was used to elute the enriched DNA, the samples was cleaned up with a MinElute purification column (QIagen) following the manufacturer instructions.

2.9.4 Verification of biotin incorporation and pull-down efficiency

Following elution, the succesfull incorporation of biotin and successful pull-down of newly replicated DNA was confirmed by spotting a small amount of the 'clicked' biotinylated DNA generated in the click reaction and the 'pulled down'-DNA on a nitrocellulose membrane and probing the membrane with an HRP-conjugated streptavidin antibody. One microliter of undiluted 'clicked' DNA was spotted along with a microliter of a 1/10 and 1/100 dilution of the clicked DNA as well as a microliter of undiluted pulled-down DNA. Biotinylated T7 primer was also spotted on the membrane in amounts ranging from 75 to 500 fmols as a standard. DNA was crosslinked to the membrane by exposure to 150 mJ UV irradiation. Following crosslinking, the membrane was briefly washed with Buffer 1 and then blocked in 3% BSA (w/v) in Buffer 1 for 30 min at 60°C. The membrane was probed with ExtrAvidin Peroxidase (Sigma Aldrich, Cat No E8386), diluted 1:10 000 in Buffer 1, for 2 h at RT with gentle agitation. Unbound antibody was removed by a wash in Buffer 1 for 15 min at RT, followed by a wash in TBS-Tween for 15 min at RT. Peroxidase (HRP) bound to the membrane was detected using an ECL detection kit as in 2.5.3. ECL was added to membrane for 1 min, after which the membrane was placed between two acetate sheets and exposed to photographic film (as in 2.5.3). Film was developed using a Konika SRX-101A developer.

2.9.5 Synthesis of complimentary DNA strands

The streptavidin enrichment protocol described in 2.9.3 requires heating to 98°C in the elution step to break the biotin-streptavidin bond and release the biotinylated DNA from the streptavidin beads. As a consequence, the DNA recovered after the elution step is single stranded. While single stranded DNA can be used a substrate for qPCR reactions, the conventional next generation sequencing library preparation protocols such as the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB, Cat No E7370) require double-stranded DNA as starting material. Therefore, to make the eluted DNA compatible with library preparation methods, complimentary strand synthesis was performed with the NEB 2nd Strand Synthesis kit (NEB, Cat No E6111S) following the manufacturer instructions. Briefly, 45 μl of the DNA isolated following streptavidin pull-down was added to 23 μl ddH_2O, 8 μl of the 10x Second Strand Synthesis Reaction Buffer, 4 μl of the Second Strand Synthesis Enzyme Mix and random hexamers. The reaction was incubated for 2.5 h at 16°C in a thermocycler and the DNA purified using a MinElute clean up column and eluted in 10 μl of ddH_2O.

2.9.6 Preparation of libraries for next generation sequencing

Library preparation for next generation sequencing was performed using the NEBNext Ultra DNA Library Prep Kit for Illumina. The volume of DNA prepared in step 2.9.5 was boosted to 55.5 μl with ddH_2O. An end repair reaction was set up, consisting of DNA, 3 μl of the End Prep Enzyme Mix (NEB) and 6.5 μl of 10 x End Repair Reaction Buffer (NEB). The reaction was incubated in a thermocycler for 30 min at 20°C, followed by 30 min at 65°C. Ligation of the NEBNext sequencing adaptor was performed by adding 15 μl of Blunt/TA Ligase Master Mix (NEB), 2.5 μl of 1:10 dilution of the NEBNext Adaptor for Illumina (NEB) and 1 μl of Ligation Enhancer. The reaction was incubated at 20°C for 15 min. To digest the hairpin adaptor, 3μl of USER enzyme was added and the reaction incubated at 37°C for 15 min. Adaptor-ligated DNA was cleaned using AMPure XP beads (Beckman Coulter) in a 1:1 ratio. 86.5 μl of AMPure XP beads were added to the reaction, mixed

thoroughly and incubated for 5 min at RT. The beads, containing the adaptor-ligated fragments were separated with a magnetic rack and supernatant was removed. The beads were washed twice with 200 µl freshly prepared ethanol for 20 s and air-dried for 5 min at RT. To elute DNA, 17 µl of 0.1 x TE were added to the beads, mixed thoroughly and incubated for 2 min at RT. Beads were again separated on a magnetic stand and the supernatant, containing the adaptor ligated DNA was collected and used in the subsequent amplification reaction. The NEB Universal Primer (NEB, Cat No E6861A, 5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC*T-3') and a sample-specific NEB Index primer (NEB, Cat No E7335S and Cat No E7500S), enabling barcoding and pooling of multiple samples for the sequencing process were used in the amplification reaction. As samples were sequenced in pools of six libraries, the manufacturer-recommended combinations of NEB Index primers 2, 5, 7, 4, 6 and 12 were used. In the amplification reaction, 15 µl of adaptor-ligated DNA fragments were mixed with 2 µl of NEB Index primer, 2µl of NEB Universal PCR Primer, 6 µl of H_2O and 25 µl of NEBNext Q5 Hot Start HiFi PCR Master Mix. Amplification was performed following the conditions in Table 2-6. Amplified DNA was then cleaned up with AMPure XP beads in a 0.9:1 beads to reaction ratio. 45 µl of AMPure beads were added to the 50 µl amplification reactions and mixed well. The beads and reaction mix was incubated at RT for 5 min and the beads, containing the amplified DNA fragments were separated on a magnetic stand. The supernatant was removed and the beads were washed with 200 µl of 80% ethanol for 2 x 30 s. The beads were air-dried for 10 min, 33 µl of 0.1 x TE were added and mixed with the beads to elute the amplified DNA. The sample was incubated at RT for 2 min before the beads were separated on a magnetic stand and 28 µl of supernatant containing the eluted DNA fragments were removed to a new tube. Amplified DNA was quantified on a Qubit (ThermoFischer Scientific) using the Qubit dsDNA HS Assay Kit (ThermoFischer Scientific, Cat No Q32851). Fragment distribution was assessed on an Agilent Bionalayser using the DNA High Sensitivity chip and corresponding kit reagents (Agilent, Cat No 5067-4626). Some samples showed a strong peak at 127 bp indicating the presence of adaptor dimers.

These samples were further purified with AMpure beads in 1.2:1 bead:sample ratio, which removed most of the primer dimers present.

Cycle Step	Temperature	Time	Cycle Number
Initial Denaturation	98°C	20 seconds	1
Denaturation Anenealing/Extension	98°C 65°C	10 seconds 75 seconds	12
Final Extension	65°C	5 min	1
Hold	4°C	Hold	

Table 2-5 Amplification conditions for Repli-Seq library preparations

2.9.7 Next generation sequencing of Repli-seq samples

Repli-seq generated libraries were sequenced by Edinburgh Genomics on an Illumina HiSeq 4000 sequencing system, generating 50 bp single-ended reads. Six barcoded libraries were pooled in equimolar amounts per single sequencing lane. Two biological replicates were sequenced for each sample.

2.9.8 Analysis of Repli-seq sequencing data

Overall assessment of the quality of the sequences was performed by FastQC (Babraham Bioinformatics). The fastq read files were then aligned to the genome using Bowtie 2 (Johns Hopkins University) and and alignment files were generated in the ". bam" format. PCR duplicate removal, file sorting and indexing was performed with Samtools 1.2 (Li et al. 2009). Read density in 1,000 bp, 10,000 bp and 100,000 bp windows across the genome were calculated using the *multicov* option in Bedtools 2.25 (Quinlan & Hall 2010). Normalised FPKM values were then generated for each window in R. Correlations between biological replicates were calculated using the corrplot R package. A single replication timing value was then calculated from the early, mid and late fraction for each window using the following equation:

$$R=0.165*FPKM_{late}/(FPKM_{early}+FPKM_{mid}+FPKM_{late})+0.495*\ FPKM_{mid}/(FPKM_{early}+FPKM_{mid}+FPKM_{late})+\ ).825*FPKM_{early}/(FPKM_{early}+FPKM_{mid}+FPKM_{late})$$

The R value was used for comparison between samples. The partitioning of the data into replication domains is described in Chapter 4.4.2.

3 Chapter 3: CFS expression, mitotic chromatin structure and transcription in RPE1 and HCT116 cells

As discussed in Chapter 1.6.2, one of the most intriguing features of CFS regions is their cell type specific expression. The differences in CFS repertoire between cell lines have been previously exploited in studies aiming to match cell type specific features with CFS fragility. Example is a study from the Debatisse lab, which compared the replication timing programme of FRA3B in lymphoblastoid cells and fibroblasts and concluded that differences in replication timing between the two were implicated in fragility (Letessier et al. 2011). To differentiate factors contributing to fragility, I compared transcription levels, replication timing and chromatin structure across active and inactive CFSs in the two different cell types. Consequently, the first stage of my study involved characterising the repertoire of CFS fragility in the two cell types and identifying sites with differential expression.

CFS break mapping is usually performed cytogenetically, by visually assigning the breaks to a chromosome band. A small number of fragile sites have also been mapped at the molecular level, by hybridisation of fluorescently-labelled BAC probes with a known genomic location, followed by assessment of the break location relative to the probe (Huang et al. 1998; Becker et al. 2002). To characterise active CFSs in the RPE1 and HCT116 cell lines, I used a combination of the two approaches – an initial cytogenetic mapping was followed by a more detailed molecular fine-mapping for a subset of the identified locations.

With the shift from a cytogenetic to a molecular approach for CFS characterisation, a transformation in the ideas about their underlying chromatin configuration has also occurred. Previously assumed to be caused exclusively by double stranded breaks, CFS-associated lesions are now suspected to be a consequence of a mixture of molecular outcomes, including concatenations and aberrant chromatin compaction in the run up to mitosis (Minocherhomji et al. 2015; Chan et al. 2009).

To determine if aberrant compaction is present at CFS locations in mitosis, I hybridised BAC probes known to cover fragile regions to metaphase chromosomes and used their signals as a probe for the underlying chromatin structure in a variety of conditions.

Finally, I focused on the contribution of transcription to CFS fragility. I compared the levels of transcription across active and inactive CFSs in the two cell lines to determine if gene expression is necessary or conducive to fragility. I also used the genome engineering CRISPR-Cas9 system to modify the transcriptional level of an active fragile site and then characterised the effect of the alteration in gene expression on the fragility and mitotic structure at that CFS.

3.1 Characterisation of CFS expression in RPE1 and HCT116 cells

As mentioned in Chapter 1.6.1, the genomic locations of expressed, or "active" CFSs differ between different cell types. Therefore, the initial step in my study involved characterisation of the patterns of CFS fragility in the two cell types I selected: the telomerase-transformed retinal pigmented epithelium cell line RPE1 and HCT116, a CIN- colorectal carcinoma-derived cell line. These two cell lines were selected as they are fast-growing, suitable for transfection and imaging studies and carry normal karyotypes. HCT116 is an ENCODE Tier 3 cell line, while the RPE1 cell line has been used for multiple studies in the Gilbert lab. Patterns of CFS fragility have never been characterised in RPE1 cells whilst CFS expression in HCT116 cells has been previously characterised (Le Tallec et al. 2013), however I chose to reassess the characterisation as some variation may be present between clonal populations of the same cell line. To perform an initial cytogenetic screen for fragile locations, I treated cells with varying concentrations of the replication-stress inducing drug aphidicolin for 24 hours and prepared metaphase chromosome spreads. I stained the chromosome spreads with the DNA minor groove binder DAPI, which results in reproducible bands across the chromosome arms. DAPI preferentially binds AT-rich regions and as a result, DAPI-bright bands are predominantly gene-poor, while GC-

rich, gene-dense regions form fainter bands. I imaged all metaphase spreads across a microscope slide for each cell type and each aphidicolin concentration. I also imaged metaphase spreads derived from control cells which were not treated with aphidicolin and did not find any CFS lesions. In the aphidicolin-treated slides, for spreads containing CFS lesions, I identified the chromosome arm harbouring the abnormality based on chromosome morphology and banding pattern. For a small number of breaks (less than 5%), it was not possible to identify the corresponding chromosomes and they were not considered in further analysis. I then identified the cytogenetic bands in which the breaks were located using two complementary approaches: a visual inference of the band according to the DAPI staining and a ratio-based approach, suggested by the NHS Lothian Clinical Cytogenetics Service. In the ratio-based approach, I measured (a) the total length, in pixels, of the chromosome arm that the break occurred on and (b) the pixel length of the distance between the centromere and the break. I then calculated (b)/ (a) and used scaled models of banded chromosomes to infer possible genomic locations for the breaks. I found that the ratios clustered along the chromosome arms, indicating recurrent breaks at CFS locations (Figure 3-1). The mid-point of each cluster was taken as a putative CFS location. However, as the fixation and spreading of chromosomes is likely to cause some distortion, molecular fine-mapping of the most frequent CFS regions was also undertaken and is described in Chapter 3.2.

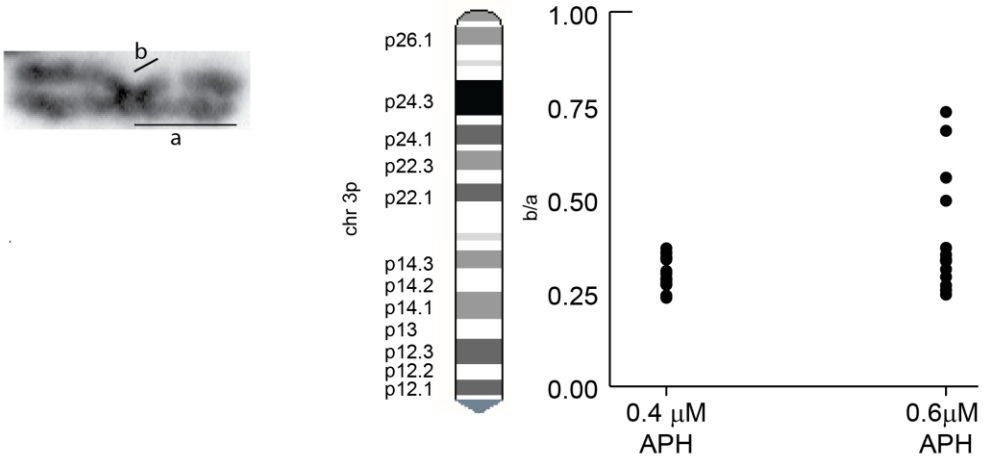


Figure 3-1: Example of ratio analysis for the localisation of breaks occurring on the p arm of chr3 following treatment with 0.4 μM and 0.6 μM aphidicolin. The length of the chromosome arm (a) and the distance from the centromere to the breakpoint point (b) were measured. The b/a ratios were plotted for all breaks occurring on Chr3p and were found to cluster, indicating recurrent lesions at 3p14.2, corresponding to the FRA3B site. Higher doses of aphidicolin revealed more fragile locations, at 3p22.1 and 3p24.2.

3.1.1 Fragile locations in RPE1 cells

The CFS repertoire in RPE1 cells was mapped after treatment with 0.4 μM aphidicolin for 24 hours. I assessed a total of 64 metaphases and found 62 breaks, resulting in a mean rate of 0.98 breaks per metaphase. Overall, I found breaks at 18 genomic locations, with the five most fragile CFSs comprising 66% of all observed breaks. While some of the fragile regions mapped to previously identified CFSs, most of them were novel CFSs specific to the RPE1 cell line. However, all of the novel fragile locations had been previously identified as sites of very rare fragility in lymphocytes (Mrasek et al. 2010). This data is summarised in Table 3-1. Various morphologies of the CFS lesions were observed, including chromatid gaps, chromosome gaps, constrictions and chromatid breaks, with chromatid gaps being

the most frequent abnormality (Figure 3-2). There was no association between particular CFS locations and lesion morphologies- defects were observed at similar proportions across all fragile genomic locations. This suggests that the processes responsible for fragility at CFS regions can give rise to a number of cytogenetic abnormalities, implying that a fundamental effect on the mitotic chromosome structure is at the root of CFS expression. When the frequency of breaks per metaphase was quantified, metaphases carrying one break were the most common and there was no indication of co-dependence of breaks: the presence of one break did not seem to increase the likelihood of a second break, suggesting CFS lesion formation is independent for each genomic location.

Loci	Associated CFS	Number of breaks observed	% of all breaks
1p31.2	FRA1C	11	18.6
3q26.32	FRA3O	10	16.9
4q32.2		7	11.9
7q21.11/7q21.12		6	10.2
2q22.2	FRA2F	5	8.5
4q31.1	FRA4C	4	6.8
13q31.1	FRA13H	3	5.1
2q34/2q35	FRA2U	2	3.4
3p24.3		2	3.4
13q14.11		1	1.7
2q31.1	FRA2G	1	1.7
7q31.2	FRA7G	1	1.7
5q31.1	FRA5C	1	1.7
17q*		1	1.7
18 centr		1	1.7
18q*		1	1.7
2p*		1	1.7
6 centr		1	1.7

Table 3-1: Fragile locations in RPE1 cells. Asterisk indicates locations where the genomic band harbouring the break could not be determined.

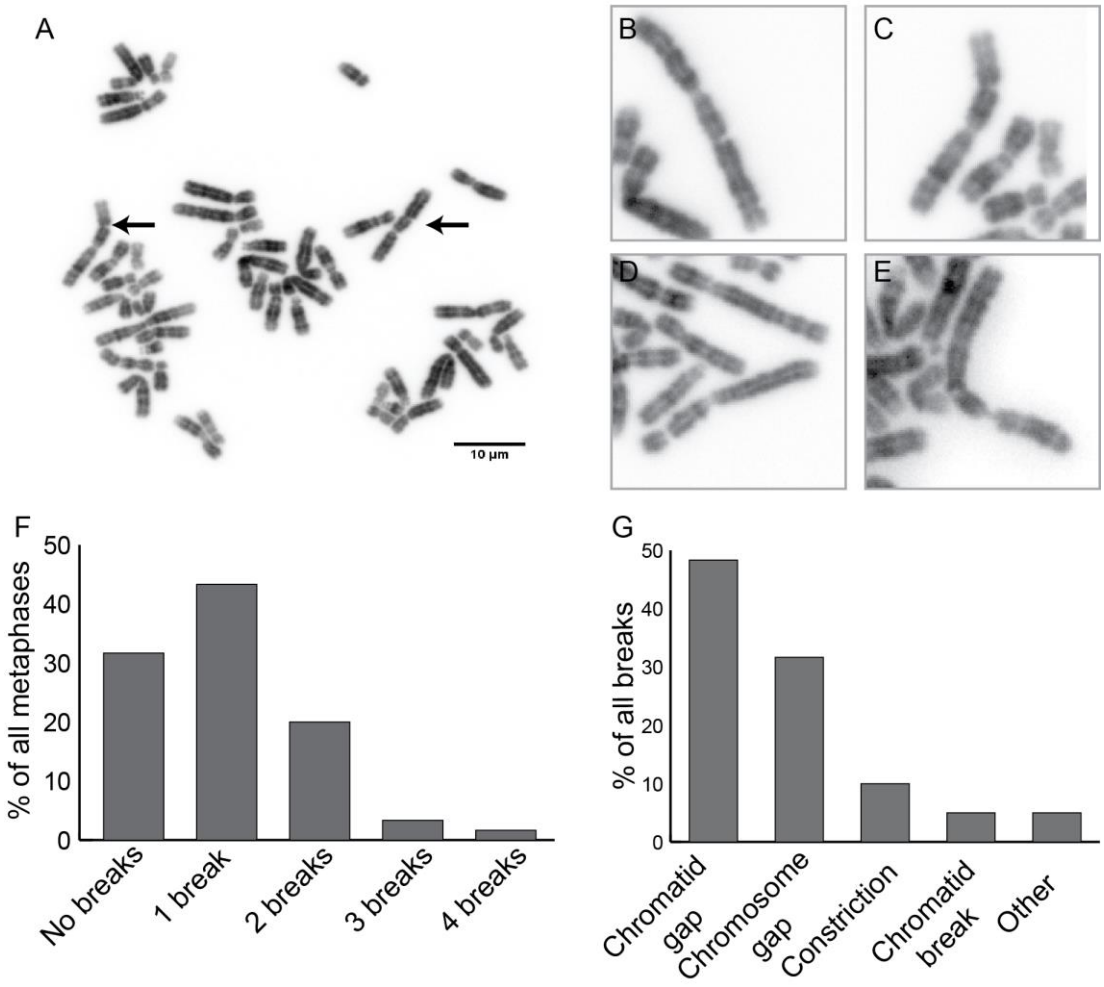


Figure 3-2 Characteristics of CFS expression in RPE1 cells. Metaphase spreads were prepared from RPE1 cells treated with 0.4 μM aphidicolin for 24 hours. The locations affected, the types of lesions produced and the numbers of breaks per metaphase were characterised. A. Representative metaphase, carrying two CFS breaks, at FRA1C and 2q22.2. B-E examples of various lesions present at CFS; B: Chromatid gap, C: Chromatid break, D: Chromosome gap, E: Constriction. F. Quantification of the frequency of metaphases carrying breaks. G. Quantification of the different defects observed at CFS upon aphidicolin treatment.

3.1.2 Fragile locations in HCT116 cells

CFS expression in HCT116 cells has been previously mapped following treatment with 0.15 μM aphidicolin for 16 hours (Le Tallec et al. 2013). Since it is unknown how CFS expression varies between different clones of the same tissue culture cell line, I performed an independent mapping of fragile locations in HCT116 cells. I treated HCT116 cells with a range of aphidicolin concentrations - 0.1, 0.2, 0.3, 0.4

and 0.6 μM aphidicolin for 24 hours and quantified the breaks I observed in metaphases derived from these cells. The number of metaphases assessed for each condition, the number of breaks observed and the rate of breaks per metaphase for the different conditions are summarised in Table 3-2. There seemed to be a general tendency for increased rates of breakages with increased aphidicolin concentrations up to 0.4 μM. At 0.6 μM, the rate of breakage appeared to be reduced, however this is due to the fact that at an aphidicolin concentration this high, a lot of metaphases appeared damaged, which impeded the identification and localisation of CFS lesions. At similar conditions of treatment as RPE1 cells – 0.4 μM APH – HCT116 cells displayed higher rates of fragility: 0.98 breaks per metaphase in RPE1 cells compared to 1.88 breaks per metaphase in HCT116 cells. This may be related to the higher levels of endogenous replication stress in the cancer-derived HCT116 cell line.

Condition	Number of metaphases	Number of breaks observed	Average breaks per metaphase
0.1 μM APH	96	28	0.29
0.2 μM APH	52	34	0.65
0.3 μM APH	59	29	0.49
0.4 μM APH	84	157	1.88
0.6 μM APH	80	124	1.55

Table 3-2 Characterisation of the CFS repertoire in HCT116 cells under different aphidicolin conditions.

In terms of break localisation, there was some variability between the different conditions, with a tendency for higher doses of aphidicolin to reveal more fragile locations. FRA3B, previously identified as the most fragile site in these cells, showed the highest frequency of breaks across the different concentrations. The most frequently identified locations and their fragility at different aphidicolin conditions are summarised in Table 3-3. Many of the sites previously identified in Le Tallec et al (2013) were also fragile in my experiments, however I failed to find breaks at two of the previously identified locations- FRA4D and FRA16B. I also identified recurrent frequent breaks at CFSs which were not previously defined as active in HCT116 cells, indicating that some variability is present between cell line clones, experimental conditions and different laboratories. These included breaks at FRA3G, FRA1C and

FRA3A. Unlike RPE1s, all of the locations I identified as fragile in HCT116 cells have been previously identified as CFSs.

Location	CFS	Identified in Le Tallec et al	% of all breaks at 0.1 μM aph	% of all breaks at 0.2 μM aph	% of all breaks at 0.3 μM aph	% of all breaks at 0.4 μM aph	% of all breaks at 0.5 μM aph
3p14.2	FRA3B	Yes	16.7	33.3	14.3	20	16.9
3p22.2	FRA3G	No	16.7	0	0	0	3.4
4q22.1	FRA4F	Yes	16.7	5.5	21.4	13.7	0
1p31.2	FRA1C	No	16.7	0	0	0	11.91
3p24.3	FRA3A	No	8.3	0	0	0	3.4
2q22.2	FRA2F	Yes	8.3	0	0	8.7	1.7
2q33.2	FRA2I	Yes	8.33	5.5	7.1	15	20.3
3q13.31	FRA3L	No	0	0	7.1	2.5	5.1
3q26.31	FRA3O	Yes	0	5.5	0	1.2	0
2q24.2	FRA2T	Yes	0	0	0	15	5.1
4q31.1	FAR4C	No	0	5.5	0	2.5	0
5q31.1	FRA5C	No	0	5.5	7.1	2.5	6.8
7q31.1	FRA7K	Yes	0	11.1	7.1	5	6.8
7q32	FRA7H	No	0	0	0	2.5	3.4

Table 3-3: CFS repertoire in HCT116 cells upon treatment with different aphidicolin concentrations.

In addition to defining active CFS locations, I also characterised the morphologies of the CFS lesions in HCT116 cells. I observed defects similar to the defects seen in RPE1 cells (chromatid gaps and breaks, chromosome gaps and constriction), but also more severe deformities. Many chromosomes carried more than one break per chromosome arm and sometimes this caused chromatid flipping (Figure 3-3). Constrictions in this cell type affected a larger region on the chromosome arm compared to RPE1 cells. In addition, the most common defect in this cell type was chromatid breaks, as opposed to RPE1s, where chromatid gaps were the most frequent. Again, there was no correspondence between particular locations and the defects observed. In addition to chromosomal lesions, defects affecting whole metaphases were seen in HCT116 cells. This included loss of sister chromatid cohesion across the metaphase and the appearance of condensed chromatin

fragments reminiscent of the “mitotic catastrophe” phenotype (Castedo et al. 2004) (Figure 3-3E).

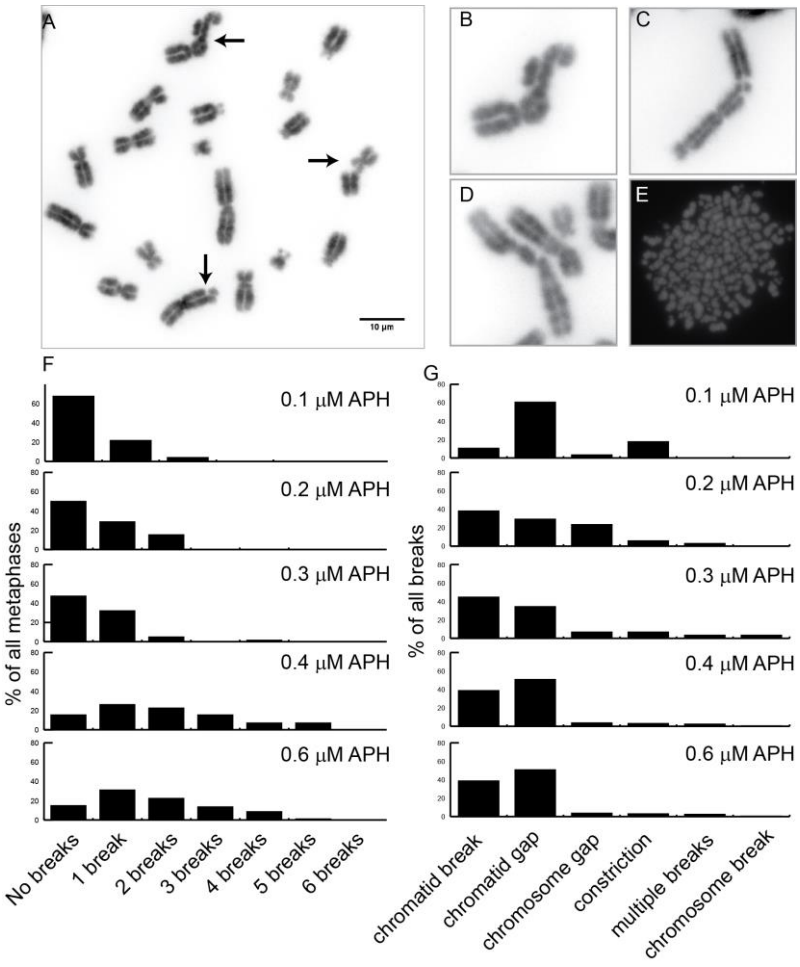


Figure 3-3 Characteristics of CFS expression in HCT116 cells. Metaphase spreads were prepared from HCT116 cells treated with 0.1, 0.2, 0.3, 0.4 and 0.6 μM aphidicolin for 24 hours. The locations affected the types of lesions produced and the numbers of breaks per metaphase were characterised. A. Part of a representative metaphase with three CFS breaks. B-E examples of extreme morphologies at CFS lesions found only in HCT116 cells; B: Chromatid flipping caused by multiple breaks, C: Chromosome carrying multiple breaks, D: Constriction affecting a large chromosome area, E: Condensed fragments indicative of mitotic catastrophe. F. Quantification of the frequency of metaphases carrying breaks. G. Quantification of the different defects observed at CFS upon aphidicolin treatment.

The characterisation of CFS expression in the two cell types indicated that most of the fragile sites were differentially expressed between RPE1 and HCT116 cells, with a very small number of sites shared between the two. As the sites with a lower

number of breaks indicated less frequent molecular events, I chose to focus only on the most highly expressed CFSs in each cell type in further experiments. For RPE1 cells, this included FRA1C, FRA3O, and the novel fragile sites at 4q32.2 and 7q21.11-7q21.12. For HCT116 cells, I focused primarily on FRA3B, FRA2I, FRA4F and FRA2T. Of the most fragile RPE1 locations, FRA1C and FRA3O showed weak fragility in HCT116 cells. FRA2F was a CFS showing some fragility in both cell types and was also included in further analyses (Figure 3-4). I anticipated that analysis of the molecular features of active CFSs and in particular CFSs shared between the two cell types would reveal which molecular processes are implicated in fragility.

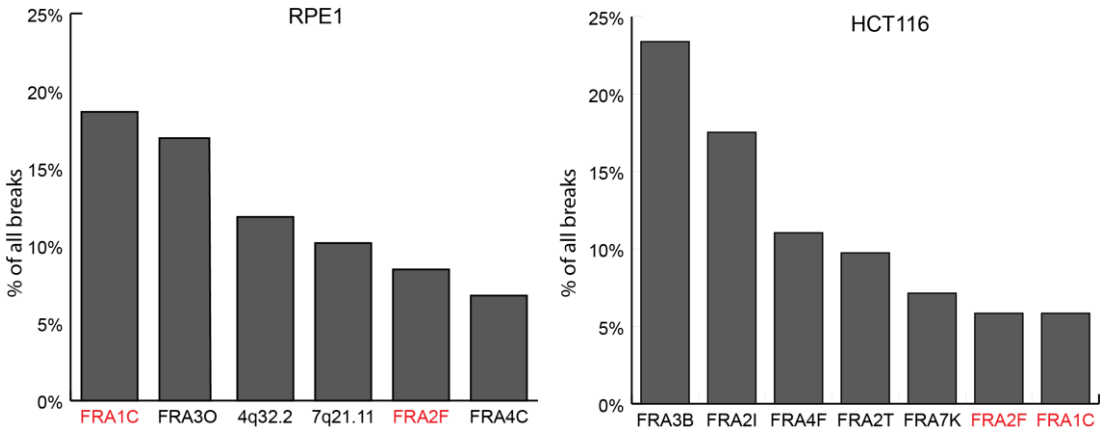


Figure 3-4 Highly expressed CFSs in RPE1 and HCT116 cell lines. CFS expression was mapped in RPE1 and HCT116 cells. Most recurrent break locations differed between the two cell lines. FRA1C and FRA2F, which were fragile in both the HCT116 and RPE1 cells, are shown in red.

3.2 Molecular mapping of CFS

Cytogenetic mapping of the CFS repertoire in RPE1 and HCT116 cells enabled localisation of CFSs to chromosome bands and previously defined fragile locations as well as a rough comparison between the CFS expression patterns in the two cell types. However, it provided very limited information on the exact genomic coordinates of the fragile regions, the sizes of the affected areas and whether any "drifting" of the breaks within the fragile regions was present. I therefore performed fine-scale molecular mapping for some of the highly expressed CFSs in the two cell types. The mapping strategy used fluorescent in-situ hybridisations (FISH) with BAC probes spanning the putative, cytogenetically identified locations to

chromosome spreads derived from cells exposed to aphidicolin. I then characterised the break locations relative to BAC positions and was able to identify the molecular coordinates of the fragile regions to the resolution of single BACs. This allowed me to assess relevant genomic features of fragile regions such as gene density, gene size and replication timing according to publicly available Repli-seq data for the IMR90 cell line.

3.2.1 Fine-mapping of CFSs in the RPE1 cell line

In the RPE1 cell line, I performed molecular mapping for two fragile locations- FRA1C and the novel CFS at 4q32.2.

3.2.1.1 FRA1C

To define the exact molecular localisation of the FRA1C CFS in RPE1 cells, I used five BAC probes spanning a 1.5 Mb region surrounding the 1p31.2 band. I performed FISH hybridisations using two or three of the BACs at a time and scored break location relative to the BAC positions (e.g. overlapping, telomeric and centromeric). In total, I analysed 265 chromosomes and found 12 breaks at FRA1C. A single BAC, RP11-482A14, located at chr1: 69399135-69576878, always overlapped with breaks (Figure 3-5). Two BACs, RP11-452B11 and RP11-44E15, neighbouring RP11-482A14 telomerically and centromerically, overlapped with 71% and 85% of breaks. However breaks frequently overlapped with these BACs only partially. BACs located telomerically from chr1:69176951 never overlapped with breaks, marking the telomeric boundary of the fragile core of FRA1C. I was unable to locate the centromeric boundary for FRA1C. However, the frequency of break overlap was dropping at the centromeric neighbour of RP11-482A14, suggesting that the fragile core of FRA1C is approximately a 0.6 Mb region covered by the three BACs. Inference of the size of individual breaks is challenging, but most breaks were found to span a region larger than RP11-482A14, which has a size of 177 kb and extend into the two neighbouring BACs, without completely occupying them. As there was some difference in the break overlap for the two neighbouring BACs it is likely that there is some minimal drift of lesions within the 0.6 Mb region defined by the three

BACs, but they are always centred on RP11-482A14. The 0.6 Mb fragile core of FRA1C is gene-poor and overlaps with a late-replicating domain in the IMR90 cells (Figure 3-5).

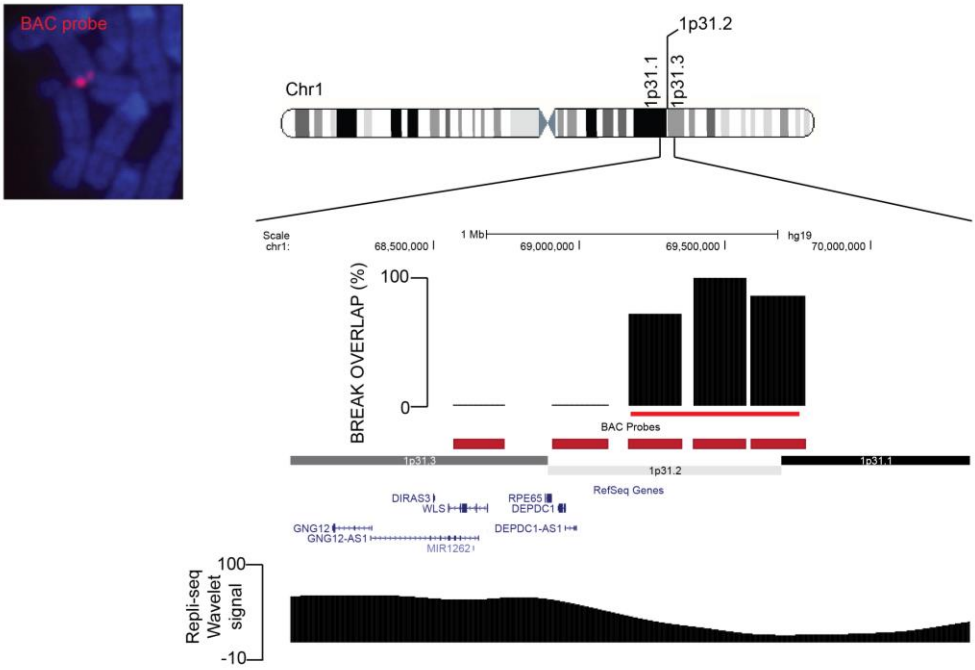


Figure 3-5 FISH-based fine-mapping of the FRA1C fragile region in RPE1 cells. BAC probes spanning a 1.5 Mb region between 1p31.1 and 1p31.3 (chr1: 68915438-69781569) were hybridised to chromosome spreads prepared from RPE1 cells treated with aphidicolin. BACs are shown in red. The locations of breaks relative to BACs were scored by how frequently the BAC overlaps with a cytogenetic break when such break is present on the chromosome. Genomic bands are shown in grey and black and RefSeq genes in blue. Bottom panel shows replication timing tracks for IMR90 cells, where higher values correspond to earlier replication timing. Inset, a representative hybridisation image showing BAC probe at 1p31.2 telomerically flanking a break on chromosome 1 (DNA counterstained in DAPI).

3.2.1.2 4q32.2

Fine-mapping of the novel fragile location at 4q32.2 was less detailed than for FRA1C. Following the cytogenetic identification of a CFS at 4q32.2, I examined the region on the UCSC Genome Browser (Kent et al. 2002) and noticed that the long gene MARCH1 was located at the boundary of 4q32.2 and 4q32.3. Long genes are frequently associated with CFSs, therefore I expected the fragile region to be located around MARCH1 and used just two BACs, RP11-946L12 (chr4: 164101959 - 164282405) and RP11-153D1 (chr4: 165720700 - 165883800), which flank the gene. I analysed 82 chromosomes and found three breaks at that site. In all cases, the two BACs partially overlapped the breaks and partially flanked them (Figure 3-6), suggesting that the fragile core of this novel CFS is contained within the 1.3 Mb region between the two BAC probes and overlapping with the MARCH1 gene (Figure 3-6). Inference of the size of breaks at this site is difficult- the two BACs are located a megabase apart and both BACS are only partially overlapped by the lesions, indicating that the cytogenetic abnormalities affect a region smaller than 1 Mb.

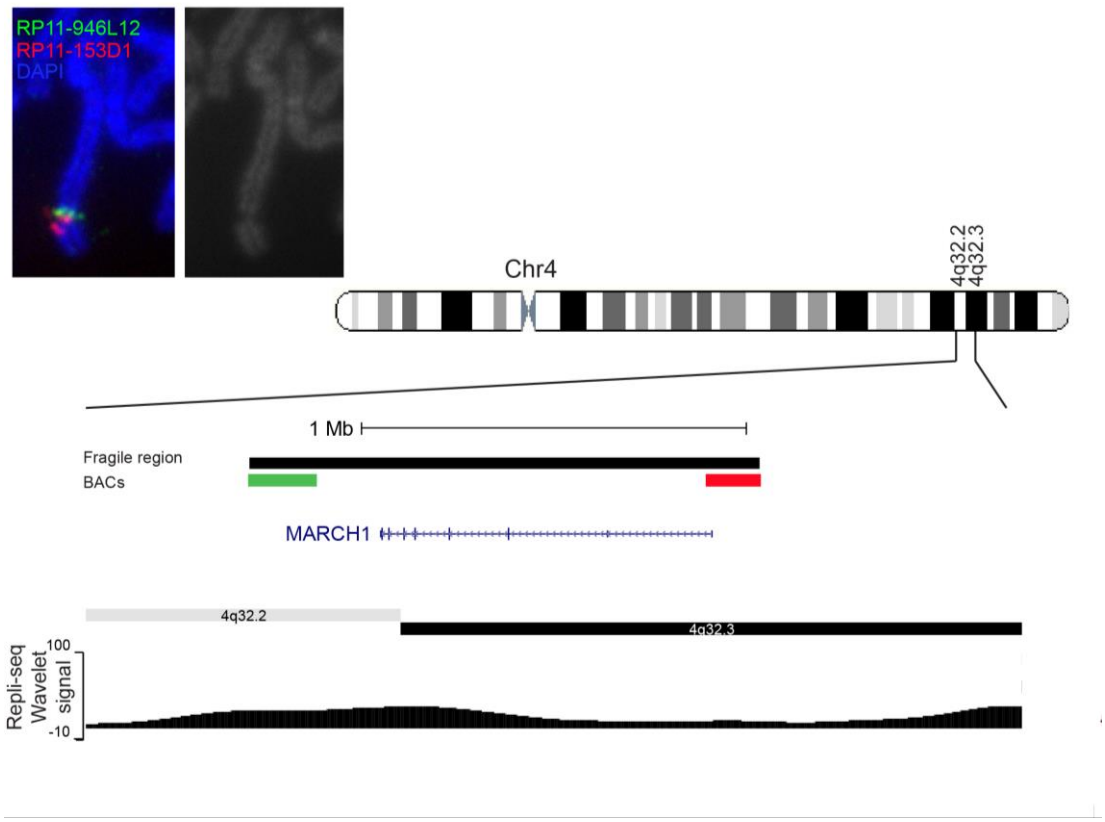


Figure 3-6 FISH-based molecular localisation of breaks at the 4q32.2 region. Two BACs surrounding the long MARCH1 gene were used to investigate if fragility at this locus coincided with the location of the gene. In the top left corner, an example of a chromosome 4 carrying a break at 4q32.2, hybridised to RP11-946L12 (green) and RP11-153D1 (red), which flank the MARCH1 gene. The two BAC probes flank the break, suggesting that fragility occurs over the MARCH1 gene body. Bottom panels: Genomic locations of the BAC probes (red and green) and the MARCH1 gene. All analysed breaks within that region appeared over the gene and coincided with a late-replicating domain in IMR90 cells. The inferred fragile region is shown in black.

3.2.2 Fine-mapping of CFSs in the HCT116 cell line

Although CFSs in HCT116 have been mapped previously, the mapping was performed cytogenetically and the precise molecular localisations of the fragile regions are unknown. I fine-mapped three sites in HCT116 cells: FRA3B, FRA4F and FRA2F.

3.2.2.1 FRA3B

FRA3B, along with FRA16D, is one of the best studied common fragile sites. It is active in lymphocytes, the colon epithelial cell line LoVo and the breast epithelial cells MCF7 and CAL-51 (Le Tallec et al. 2013; Wang et al. 1999). The molecular localisation of cytogenetic lesions at FRA3B in lymphocytes has been the subject of a number of studies; the most conclusive characterisation was performed by hybridisation with BAC probes spanning the region and positioned the breaks to a 4 Mb region spanning from 3p14.1 to the FHIT gene at 3p14.2 (Becker et al. 2002). However, the localisation of breaks in HCT116 cells is unknown. I mapped the break positions using three fosmid probes, spaced 1 Mb apart across a 3 Mb region centred on the FHIT gene (Figure 3-7). I scored the position of 18 breaks from 62 metaphases relative to the fosmids. In my analysis, all breaks appeared within a 1 Mb region between two of the three probes, which I called F3B1 (chr3: 59423151-59459364) and F3B2 (chr3: 60950550- 60993172). F3B1 flanks FHIT telomerically while F3B2 is near the centromeric end of the gene (Figure 3-7). F3B1 and F3B2 flanked the breaks in most cases and were occasionally partly inside the breaks, suggesting some minimal drifting of the lesions within the 1 Mb region. As previously described for FRA3B, the whole region coincides with a late-replicating domain, defined by publicly available replication timing data.

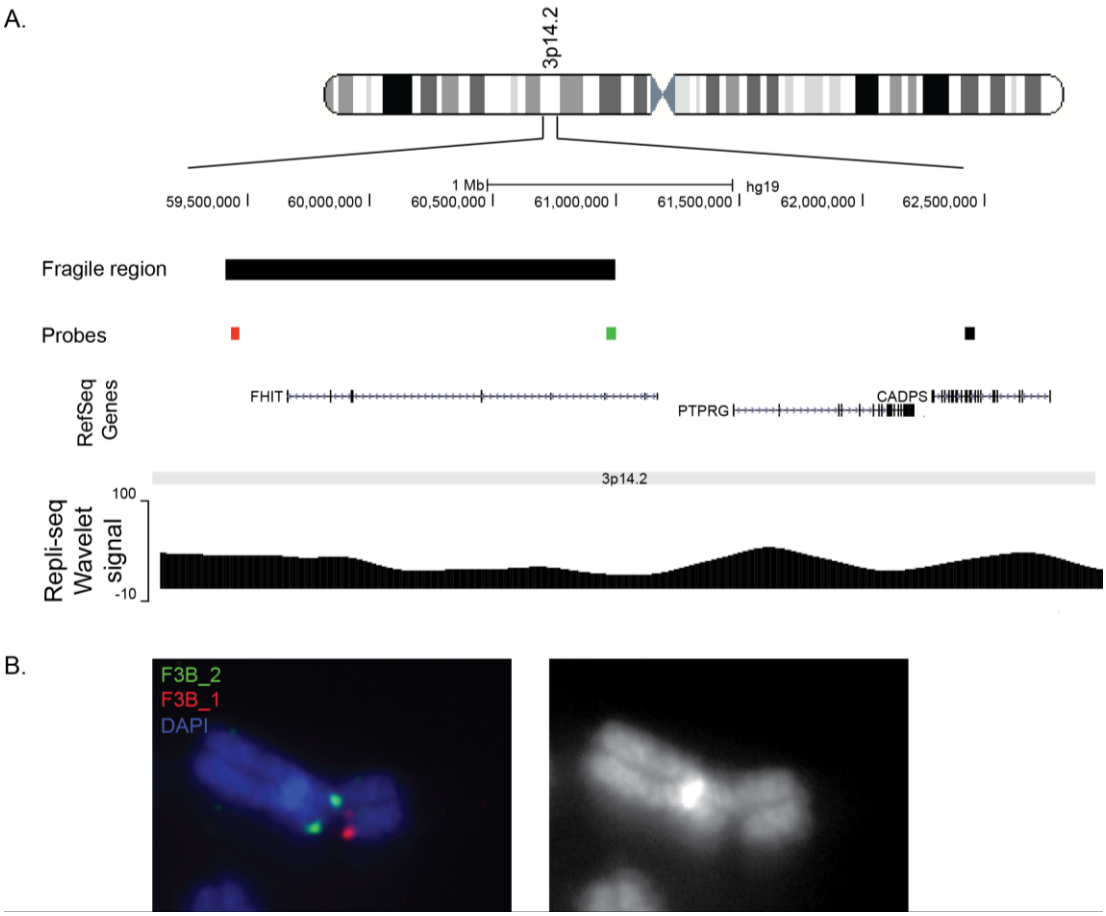


Figure 3-7 FISH-based molecular localisation of breaks at the FRA3B site. A. Three probes, spanning a 3 Mb region (chr3: 59423151-62450319) centred on FHIT were used to map break positions at the FRA3B locus. All breaks localised between the F3B1 (red) and F3B2 (green) probes. The fragile region (black, top panel) corresponded to a late replicating domain (bottom panel). B. Representative image of chromosome 3 carrying a break at FRA3B hybridised to F3B1 (red) and F3B2 (green), which flank the break. Chromosome is counter-stained with DAPI.

3.2.2.2 FRA4F

The molecular localisation of lesions at the FRA4F CFS was mapped via FISH hybridisations with seven BAC probes spanning a 10 Mb region surrounding the 4q22.1 – 4q22.2 boundary and including two large genes: CCSER1 and GRID3. A total of 342 chromosomes were analysed for that locus and they contained 28 breaks. All breaks overlapped with probe RP11-351L22, located at

chr4: 92912674-93073948, towards the telomeric end of GRID2 (Figure 3-8). Most breaks overlapped with probes located both centromerically and telomerically from RP11-351L22, but not completely, suggesting some drift of lesions, or variability of their extent, within a 5 Mb region. The telomeric boundary of the fragile regions was marked by probe RP11-688G4 (chr4: 96196712- 96376784) at 4q22.3. Centromerically, very few lesions extended beyond probe RP11-1053C2 (chr4: 89213170-89389643), located in 4q22.1, centromerically from the CCSER1 gene body. Therefore, fragility at FRA4F extended along a 5 Mb region, making it larger than all other sites mapped within this project. This site also overlaps a late-replicating domain.

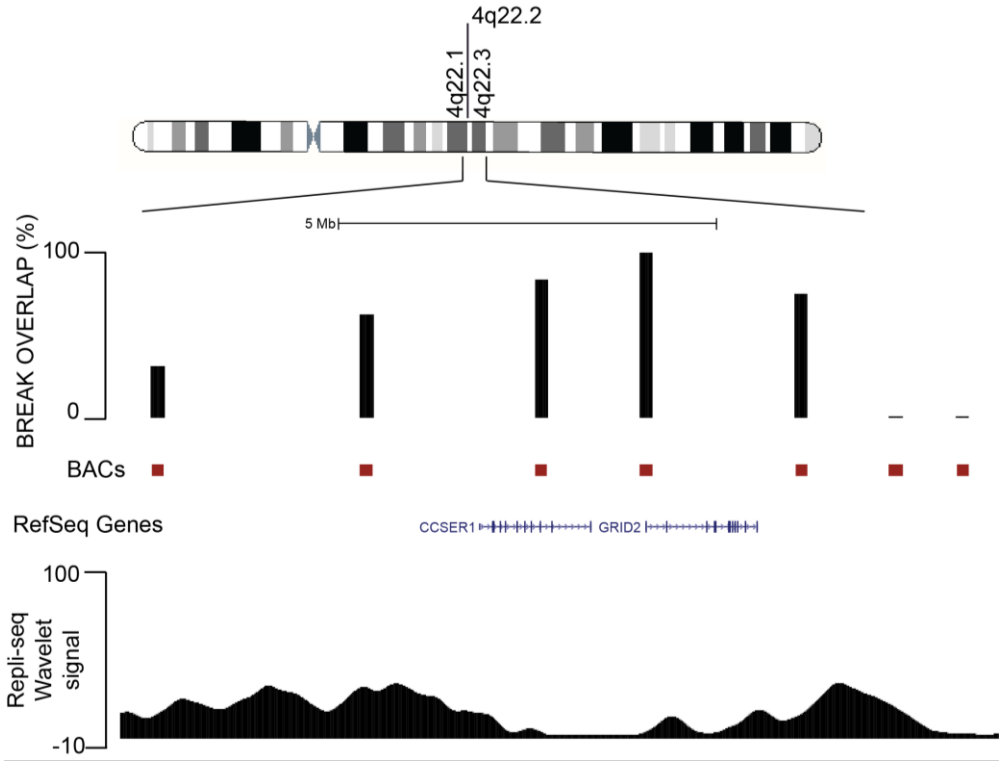


Figure 3-8 FISH-based molecular localisation of breaks at FRA4F. Fine-mapping of this site was performed with seven BAC probes spanning a 10 Mb region at chr6: 87485968- 96376784 (shown in red). Top panel shows break overlap for each BAC probe across the region. RP11-351L22, located at the 5' end of the GRID2 gene (blue) was overlapped by all breaks within the region. Breaks also frequently overlapped with probes surrounding RP11-351L22 within a 5 Mb region. The 5 Mb fragile region overlaps with a late-replicating domain in the IMR90 cell line (bottom panel).

1.1.7.1 FRA2F

The final fragile location I fine-mapped was the FRA2F locus at the 2q22.1 – 2q22.2 boundary. This location contains the LRP1B gene, which is nearly 2 Mb long and sits at the boundary between the two cytogenetic bands. I expected fragility would occur in the vicinity of the gene and selected four BAC probes within a 5 Mb region centred on the gene for the fine-mapping. 126 chromosomes were analysed of which 19 carried a break at the FRA2F CFS. All of the breaks occurred telomerically from RP11-436I1 (chr2: 136818800-137001238) and RP11-236P10 (chr2: 141182186-141337859), the two probes within 2q22.1. RP11-15G12 (chr2: 142948672-143108539) and RP11-56K5 (chr2: 145006565-145166970) overlapped with breaks. Breaks sometimes extended beyond the two probes and were sometimes flanked telomerically by them, indicating some drift of the telomeric boundary of this region. The localisation of the breaks around this region is surprising – they appear telomerically from the LRP1B gene body and just outside the late replication timing domain which LRP1B spans.

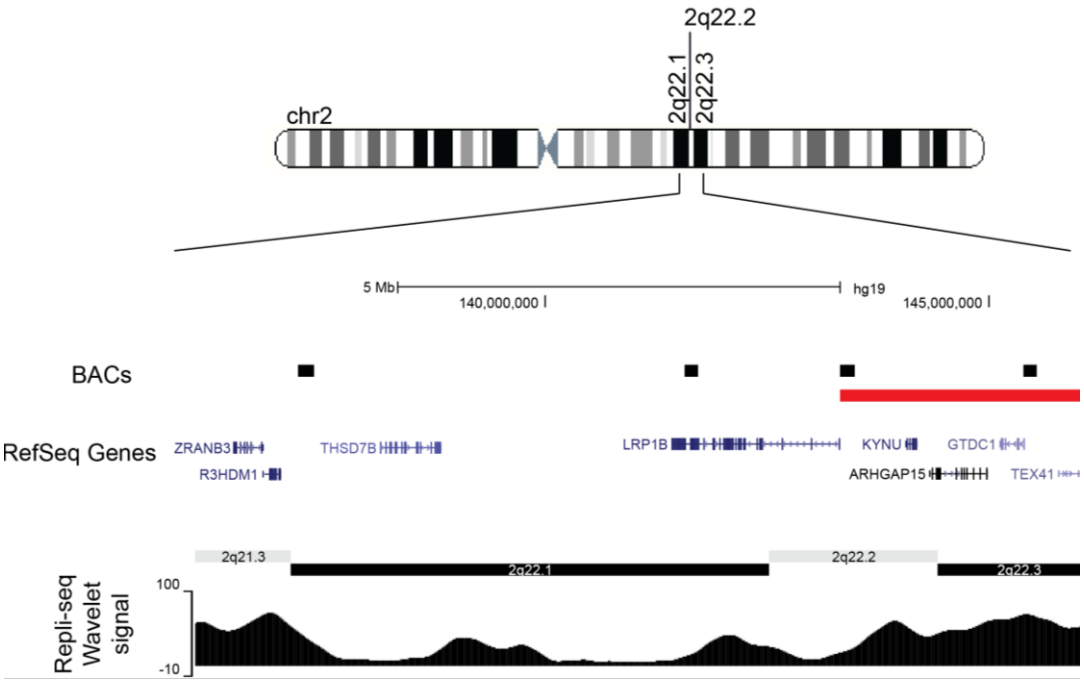


Figure 3-9: FISH-based molecular localisation of breaks at the FRA2F site. Four BAC probes over a 5 Mb region were used to map the localisation of breaks at the FRA2F region. Surprisingly, breaks were found to cluster at the telomeric end of that region, at the 2q22.2 -2q22.3 boundary and outside the LRP1B gene body and the late replication timing domain (defined from publicly available data from the IMR90 cell line) that encompasses it. The location of breaks is indicated by the red bar.

3.3 Investigating mitotic chromatin structure at CFS

An unresolved question about common fragile sites is related to the nature of the molecular structures underlying the metaphase cytogenetic lesions. Historically, these lesions were assumed to represent single-stranded or double-stranded DNA breaks. However, as the interplay between the processes of DNA replication and mitotic compaction becomes clearer, the idea that CFS lesions may represent problems with mitotic condensation, rather than breaks, has become more popular. It has been demonstrated that condensin binding and the subsequent processes of chromosome condensation and sister chromatid separation are dependent on successful and timely DNA replication (Ono et al. 2013). CFS regions are likely to experience replication delays or even remain unreplicated in G2, and it is easy to imagine how this may reverberate through subsequent steps of mitotic chromatin

assembly. Therefore, I set out to investigate the chromatin structures of active CFS regions on metaphase chromosomes using FISH. I focused on two CFS regions: FRA1C and FRA4F, which I assayed in RPE1 and HCT116 cells, respectively. I used FISH-based hybridisations with BAC probes validated in the fine-mapping experiments as a tool to examine chromatin state in the two CFSs in chromosomes derived from control cells and from cells treated with aphidicolin.

3.3.1 Mitotic chromatin across CFS

Throughout my fine-mapping experiments, I was able to observe interesting features of chromatin compaction at common fragile sites. At some of the breaks, I observed that FISH probe signals appeared within the breaks, suggesting that there might be DNA present within the site. I quantified the fluorescence intensity for probes at two break sites – FRA1C and FRA4F, on chromosomes with cytological abnormalities. I found that the fluorescence intensity of the probes peaked over the DAPI – faint regions marking the CFS, consistent with DNA being present within the cytigenetically visible breaks (Figure 3-10).

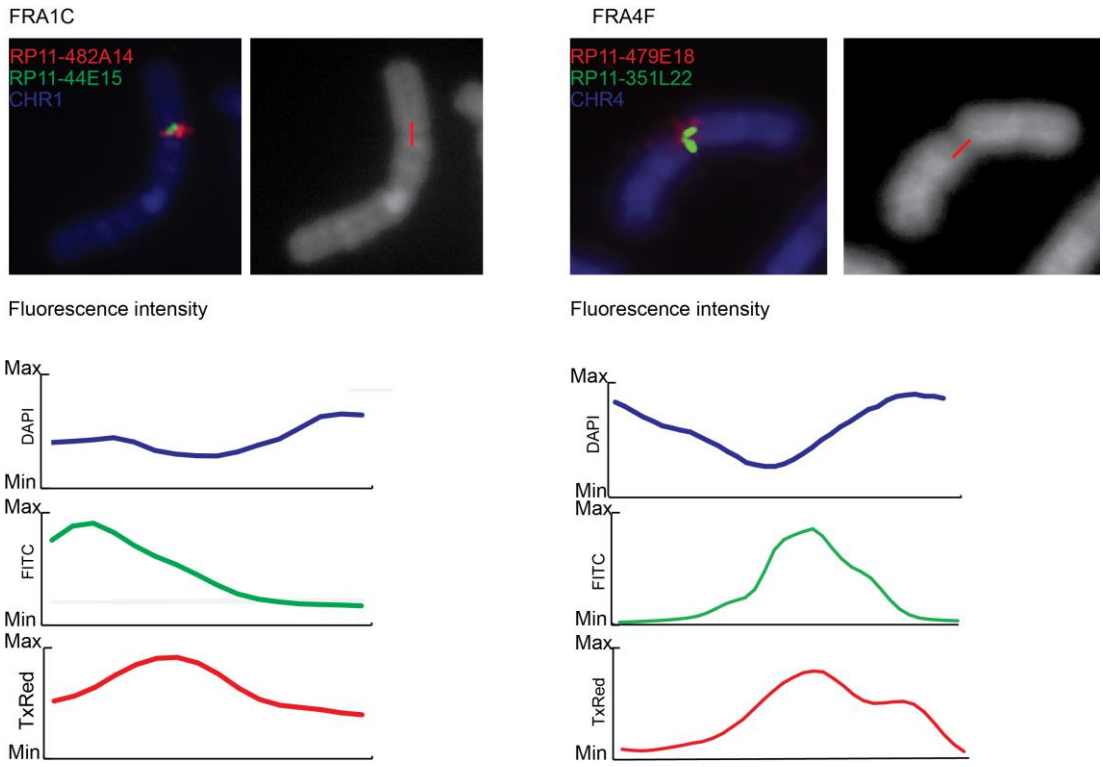


Figure 3-10 Fluorescence intensity of FISH probes spanning CFS breaks. I quantified the fluorescence intensity over breaks at the FRA1C (left) and FRA4F (right) CFSs. Top panel shows FISH images for the two sites and bottom panel shows the fluorescent intensities for the region marked in red in the images. High intensity signals appear over DAPI-faint areas at both sites.

In addition to chromosomes carrying cytogenetic breaks, I also observed the chromatin states, as identified by the BAC probes, on chromosomes which did not carry any obvious abnormalities. Curiously, BAC probes hybridising to fragile loci frequently displayed atypical signals: instead of the symmetrical signals frequently observed on mitotic chromosomes, signals at CFSs would appear asymmetric, fragmented, concatenated between two chromatids and in the most extreme cases, appear to extend outside of the chromosome scaffold, reminiscent of uncompacted chromosome loops (Figure 3-11). Curiously, a similar phenotype has been described at telomeres previously, resulting from replication stress and depletion of components of the shelterin complex (Sfeir et al. 2009). I did not find any correspondence between the types of signals present and different CFS locations. However, I hypothesised that these signals indicated problems with the underlying chromosome structure, suggesting that even on cytogenetically normal

chromosomes, CFS regions show abnormal mitotic compaction. I therefore quantified the frequency of these atypical signals for two CFSs: FRA1C in RPE1 cells and FRA4F in HCT116 cells.

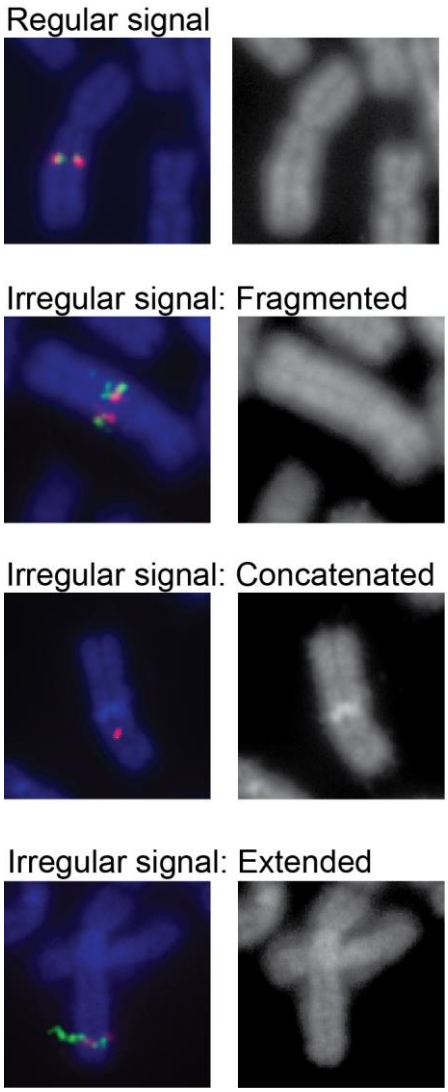


Figure 3-11 Atypical FISH probe signals at CFS loci. From top to bottom: regular signals showing symmetrical spots on the two chromatids; fragmented: signals appear "fragmented" and are formed by multiple spots; concatenated: a single spot between two chromatids; extended –signal extends from chromosome scaffold, reminiscent of an uncompacted loop.

3.3.2 FRA4F

I quantified the frequency of irregular mitotic signals for six of the seven BAC probes used to fine-map FRA4F, only including cytogenetically normal chromosomes in my analysis. I performed the scoring in chromosomes derived from cells that have

been treated with 0.4 μM APH for 24 h and scored the BAC signal as either “regular” or “irregular”. The regular group included symmetrical signals made up of two spherical spots, while the irregular group encompassed all asymmetric, concatenated, fragmented and extended signals. Interesingly, the distribution of the irregular signals across the BACs correlated to the frequency of break overlap: the highest frequency of irregular signals was found around BAC probes which most frequently overlapped breaks at FRA4F, and tailed off at BACs which rarely overlapped with breaks (Figure 3-12). Similarly to the break distribution at this site, the frequency of atypical signals peaked at the centromeric boundary of the GRID2 gene body and within the late replicating domain overlapping it. If the irregular signals are a sign of chromatin misfolding, this result suggests that the cytogenetic abnormalities at the FRA4F CFS site arise within a region that is highly prone to incorrect mitotic compaction in the presence of replication stress.

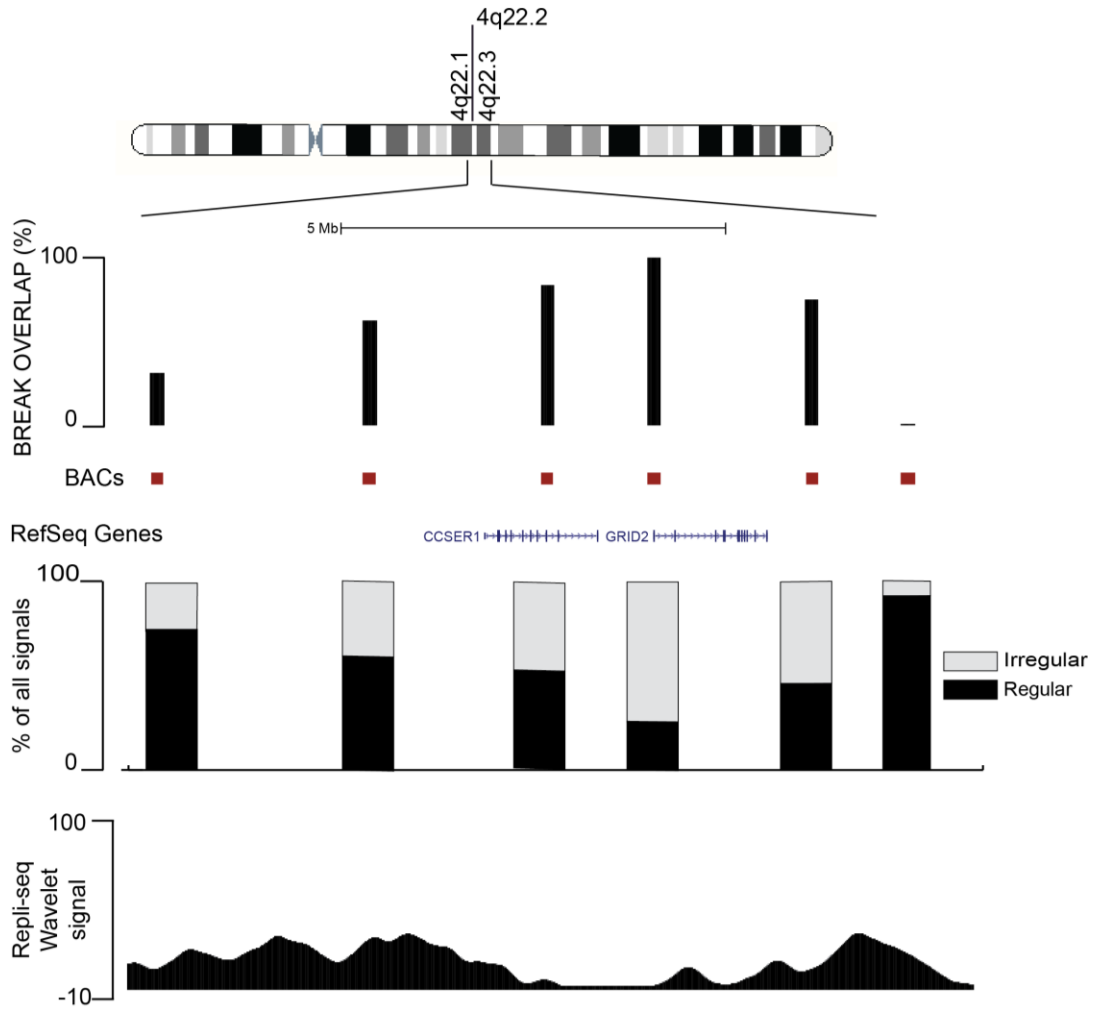


Figure 3-12 Atypical FISH probe signals across the FRA4F locus. Quantification of the frequency of atypical signals across six BAC probes used for fine-mapping FRA4F (red). Top panel shows the lesion overlap for each probe, while the mid panel shows the frequency of irregular signasl at each BAC. Bottom panel shows replication timing data for the IMR-90 cell line.

3.3.3 FRA1C

I next wanted to determine whether the observations for FRA4F could be confirmed for a different fragile site and in a different cell line. I therefore quantified the frequency of irregular signals across the FRA1C site in RPE1 cells following treatment with 0.4 μM APH for 24 hours. Again, I scored BAC signals using the same criteria as for FRA4F. Although the frequency of atypical signals was lower in the RPE1 cell line, I observed a similar trend as for FRA4F: the highest proportion of atypical signals was seen for the BAC probes where breaks were most frequently

found. The rates of atypical signals were then reduced towards the boundary of the fragile region, suggesting that there is an overlap between the area of mitotic mis-folding and the fragile core of the site. Unlike the FRA4F site, no long genes are present within FRA1C, although like most fragile sites it overlaps a region of late replication. This indicates that altered chromosome folding accompanying fragility at these two CFSs are independent of site-specific features such as long genes.

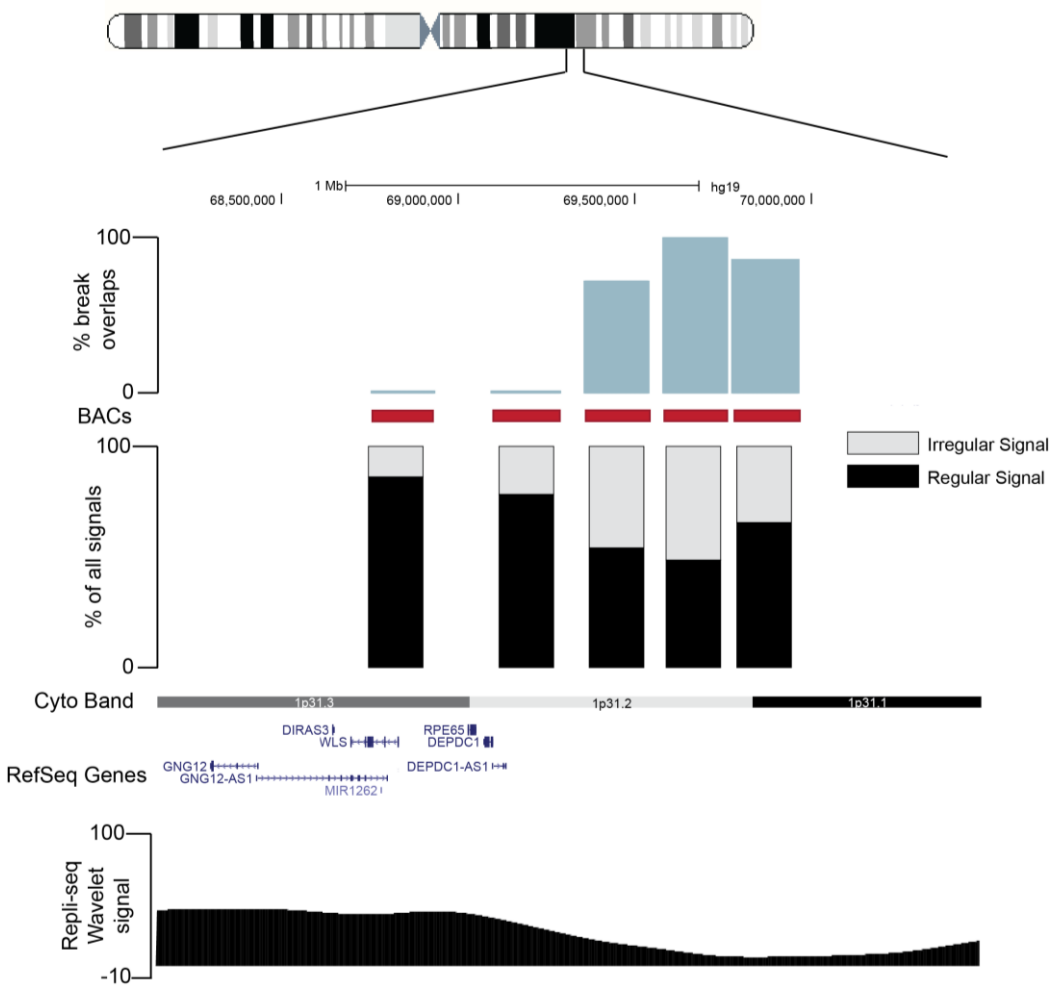


Figure 3-13 Atypical FISH probe signals across the FRA1C locus. I quantified the frequency of atypical signals across five of the BAC probes used for fine-mapping FRA1C (shown in red). Top panel shows the lesion overlap for each probe, while the mid panel shows the frequency of irregular signal at each BAC. Bottom panel shows replication timing data for the IMR-90 cell line. Similar to the case at FRA4F, frequency of irregular signals at BAC probes mirrors the prevalence of break overlap for this site.

This data indicates that FRA4F and FRA1C both overlap with regions of localised mitotic mis-compaction in HCT116 and RPE1 cells respectively. It is highly likely that such mis-compaction mechanistically contributes to the generation of mitotic lesions at these two sites. I therefore set out to investigate how replication stress contributes to mitotic folding problems at these two loci.

3.3.4 Influence of replication stress on mitotic compaction at FRA4F and FRA1C

Initially, I set out to assess the frequency of mis-folding signals in the two locations in the absence of aphidicolin-generated replication stress. I performed hybridisations with the probes showing the highest frequencies of mis-folding in the FRA1C and FRA4F regions in chromosomes derived from cells which were not exposed to aphidicolin. In addition, I also performed hybridisations with two control BAC probes localising to non-fragile locations in the genome and assessed the frequency of regular and irregular signals at these probes. For both of the CFS probes, I found a significant increase in the frequency of irregular signals when cells were treated with aphidicolin (Figure 3-14). For FRA1C, the proportion of regular signals decreased from 70% in the absence of aphidicolin to 53 % upon induction of replication stress (χ-squared test p-value = 0.001). At FRA4F, the proportion of regular signals decreased from 45% to 26% upon aphidicolin treatment (χ-squared p-value = 0.01). However even in the absence of aphidicolin, some mis-folding was present in these regions, especially at the FRA4F locus in HCT116 cells: 51 % of signals at that site appeared irregular, which may be a reflection of the endogenous replication stress present in cancer-derived cells. The high rates of misfolding at FRA1C and FRA4F in untreated cells may indicate that CFS regions suffer problems with mitotic compaction even when cells go through an unperturbed replication. In contrast, the control probes showed very low level of mis-folding in untreated cells (6.8 % in RPE1 cells and 10.9 % in HCT116 cells). Surprisingly, these probes also showed a significant increase in the frequency of irregular signals upon aphidicolin treatment, despite the fact that they never overlapped with mitotic lesions.

However even after aphidicolin treatment, the rates of mitotic misfolding at the control probes remained substantially lower than at CFS regions. These results indicated that replication stress interferes with mitotic compaction, particularly at CFS regions.

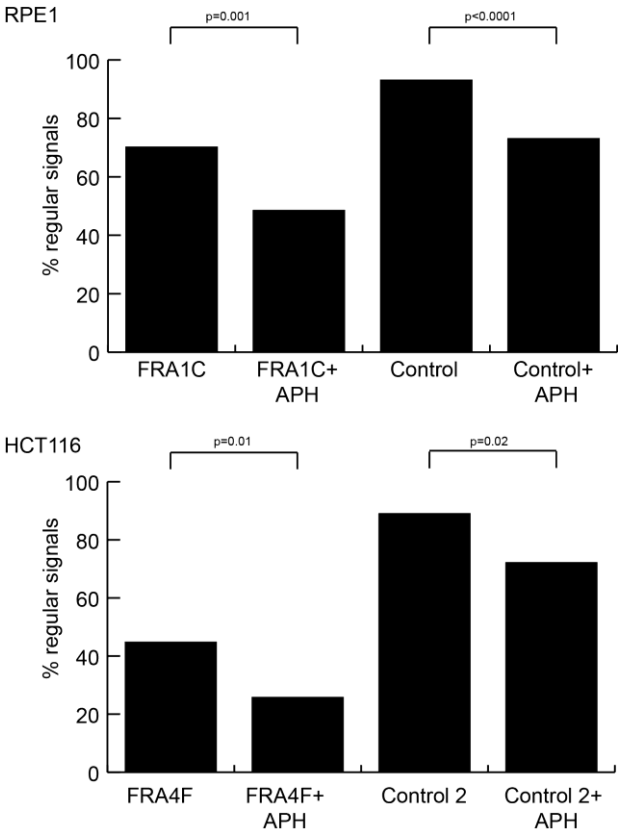


Figure 3-14 Effect of aphidicolin on mis-compaction in mitosis. Chromosome spreads from aphidicolin-treated and control RPE1 and HCT116 cells were hybridised to probes for FRA1C (RPE1 cells), FRA4F (HCT116 cells) and control BACs mapped to non-fragile regions. Frequency of regular signals for each locus under each condition was calculated. Top graph summarises results from RPE1 cells. Both the control probe and the FRA1C probe showed a significance increase in misfolding upon aphidicolin treatment, however rates of misfolding at FRA1C were much higher compared to the control. A similar trend was observed in HCT116 cells (bottom panel). FRA4F and a probe hybridising to a non-fragile genomic location both showed an increase in mis-folding signals upon replication stress induction. Rates of mis-folding were substantially higher at FRA4F compared to the control region.

3.3.5 Investigating the process of mitotic compaction at CFS regions

I next aimed to investigate how the process of chromatin folding for mitosis differs at CFS regions compared to non-fragile regions. To do this I used the drug calyculin, a protein phosphatase inhibitor known to trigger chromosome compaction whatever the cell cycle stage - a process known as premature chromosome condensation (PCC). A remarkable property of PCC – derived chromosomes is that their morphology is indicative of the cell cycle stage they originated from (Kanda et al. 1999). G1-derived chromosomes have a single, zig-zag shaped chromatid (Figure 3-15). S-phase chromosomes do not have the ability to compact completely and instead form condensed fragments, resulting in a "pulverized" appearance. Late S and G2 cells form chromosomes which are morphologically very similar to their mitotic counterparts, but are longer with a "fuzzy" appearance. Mitotic chromosomes from calyculin – treated cells appear normal and cannot be distinguished from chromosomes derived from untreated mitotic cells, although an increased frequency of breaks at CFSs has been observed following calyculin treatment (El Achkar et al. 2005). To test the ability of a CFS region and a non-fragile region on chromosome 11 to compact at different stages of the cell cycle, I treated HCT116 cells with 50 ng/ml calyculin for 1 hour to induce PCC. Prior to calyculin treatment, I pulsed cells with EdU for six hours, which allowed visualisation of replicated regions and aided the morphological categorisation of chromosomes into the different stages of the cell cycle: G1 chromosomes were EdU-negative, while S and G2 chromosomes were EdU positive. I prepared chromosome spreads from the calyculin –treated, EdU-labelled cells and hybridised them to RP11-351L22, the BAC probe showing the highest frequency of mis-folding at the FRA4F locus as well as a control BAC hybridising to chromosome 11, which never overlapped with breaks. Following FISH, a click-chemistry based staining was also performed on the slides to allow visualisation of EdU – positive cells. I quantified the frequency of uncompacted, "extended" signals at each of the two probes and throughout different cell cycle stages (Figure 3-15). Interestingly, I

found that these “extended” signals had similar prevalence at the two locations during the G1 and S phases in the cell cycle. However, as cells transitioned into G2 and M, the frequency of uncompacted signals sharply dropped off at the chromosome 11 locus. In contrast, at FRA4F, extended signals were present with comparable frequency throughout the cell cycle, indicating that the process which allows non-fragile locations to re-set their chromatin environment and compact for mitosis may be disrupted at CFSs such as FRA4F.

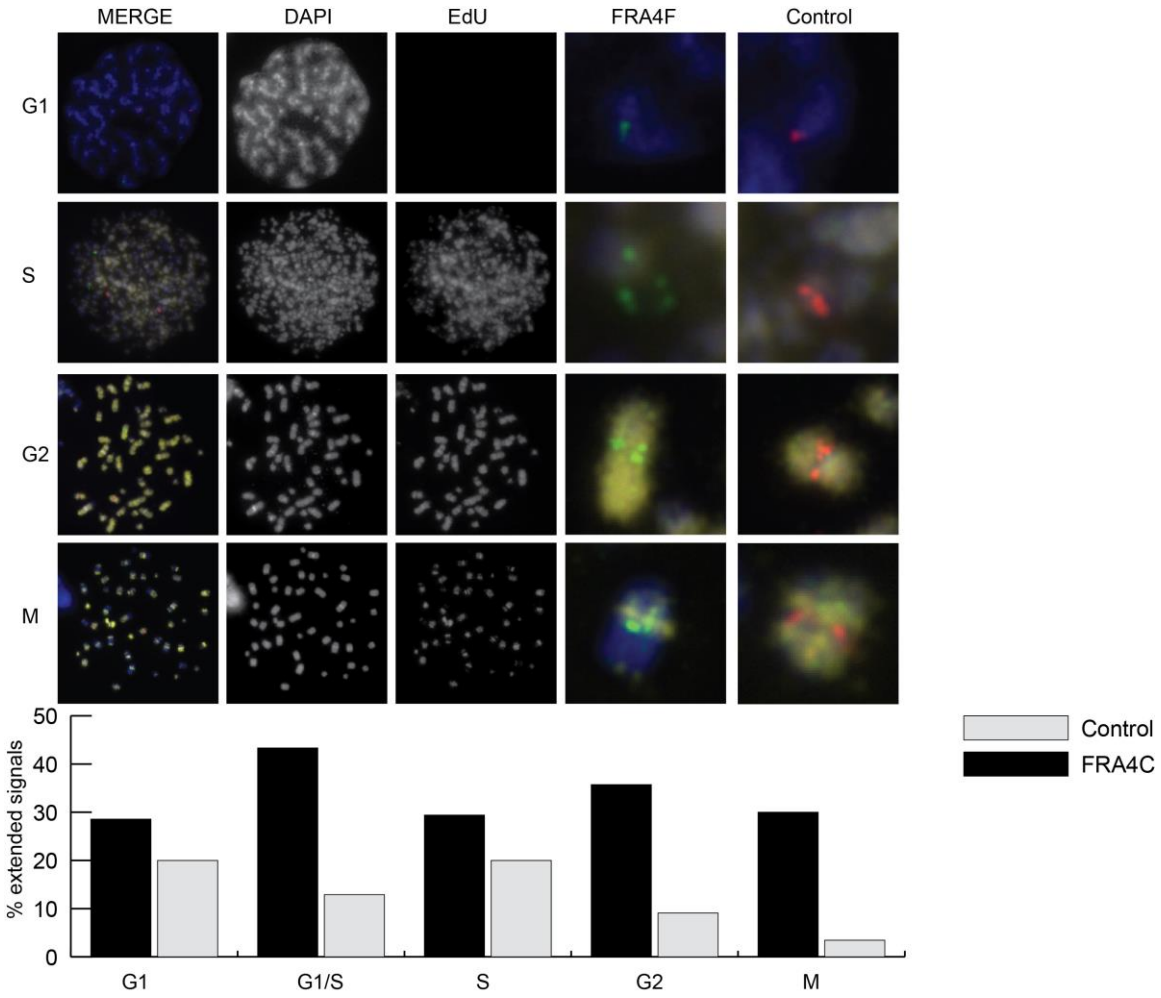


Figure 3-15 Premature chromosome condensation reveals differential compaction of CFSs and non-fragile sites prior to mitosis. HCT116 cells were pulsed with the thymidine analogue EdU, then treated with calyculin to induce premature chromosome condensation. Chromosomes from these cells were hybridised to probes for FRA4F CFS (green) and a control, non – fragile locus (red). PCC results in different chromosome morphologies, dependent on the cell cycle stage the cells were derived from which are shown in the top panel: G1 chromosomes have a single chromatid following a zig-zag path and are EdU negative. S-phase chromosomes have a "pulverised" appearance and have diffuse EdU staining. G2 chromosomes appear similar to mitotic chromosomes, but are longer and "fuzzier" and are also EdU positive. Mitotic chromosomes appear normal and only the late replicating bands were labelled with EdU. Bottom panel shows the quantification of extended, uncompacted signal for the two FISH probes across different cell cycle stages.

To determine if FISH signals could be used as a probe for chromatin state, I also quantified the frequencies of single spot signals and signals consisting of two spots

or more at each cell cycle stage. I observed that as cells moved through the cell cycle, the frequency of single spot signals was decreased and the frequency of signals consisting of two or more spots was increased for both the FRA4F and the control probe. This indicated that FISH signals can provide a reliable readout for the underlying chromatin state of a locus and is consistent with previous studies which have used the number of separate signals at a locus as an indication of whether the site has replicated (Palakodeti et al. 2009).

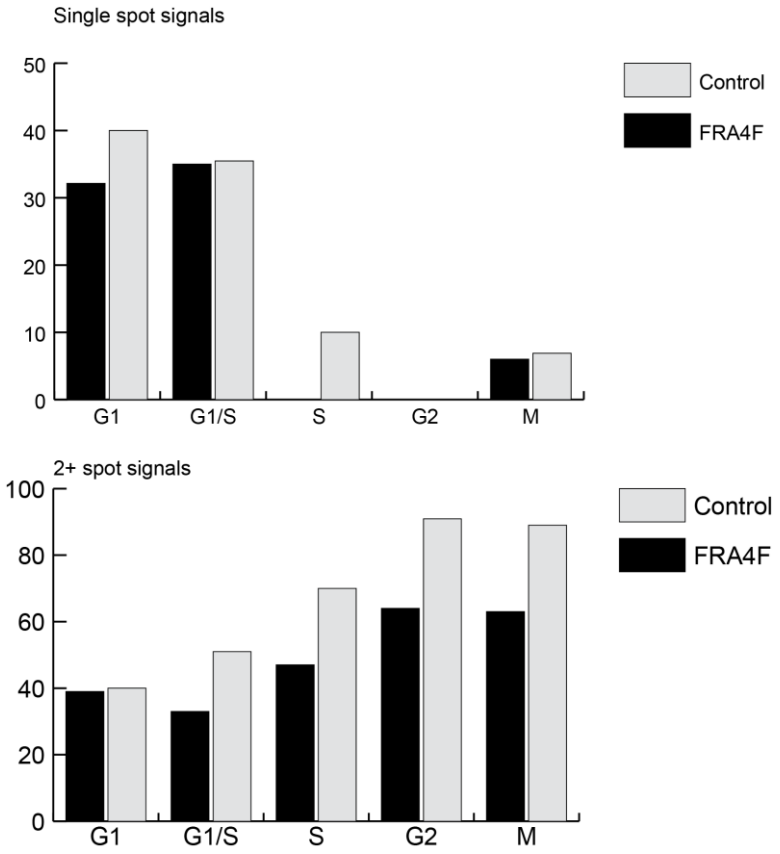


Figure 3-16 Number of separate signals at a FISH probe is indicative of the replication status of the locus. Chromosomes were prepared from calyculin treated cells and hybridised to probes for the FRA4F locus and a control locus. The frequency of probes producing single spot signals (top graph) and two or more signals (bottom graph) was characterised for different stages of the cell cycle. The frequency of single spot signals decreased for both probes as cells transitioned through S-phase. The frequency of two or more signals per probe increased throughout the cell cycle, indicating the loci have undergone replication.

3.4 Transcription at CFS

The question of whether transcription influences the fragility of CFS regions is unresolved. Contribution from transcription is suspected, as a high proportion of CFSs are located in the vicinity, or span, long genes. The difficulty in correlating gene expression and fragility derives from the cell-type specific nature of these two factors- not many studies have measured CFS fragility and gene expression in the same cell type. A study in HCT116 measured expression levels across 24 CFS – associated genes and found no signs of relationship between transcription and fragility in the presence or absence of aphidicolin (Le Tallec et al. 2013). In contrast, a more mechanistic study found that transcribing the full length of long genes such as the CFS – associated FHIT and WWOX can take more than the entire cell cycle, which would suggest that transcription and replication must happen concurrently at these loci (Helmrich et al. 2011). This study also found a surprisingly good correlation between gene expression of very long genes and metaphase CFS breaks in lymphoblast and myoblast cells. Furthermore, the authors demonstrated the formation of R-loops at FRA3B and an increase in CFS fragility in the absence of RNase H, concluding that RNA: DNA hybrids are implicated in CFS breakage. CFSs have also been identified as regions where mutations accumulate as a result of transcriptional stress upon RECQL5 depletion (Saponaro et al. 2014).

I wanted to investigate the effect of transcription across the set of CFSs I have identified in the RPE1 and HCT116 cell lines. As this included CFSs with differential fragility in the two cell lines, I reasoned that comparing the transcriptional landscapes across a number of fragile regions on stable and unstable backgrounds would be very informative. I therefore performed total, ribosome-depleted RNA sequencing in the RPE1 cell line and analysed a publicly available dataset of RNA-seq for the HCT116 cell line.

RNA-seq in the RPE1 cell line was performed as described in 2.4.3. A total of 26,476,759 50-bp reads were analysed, of which 78.73% mapped to ribosomal DNA genes and were removed from further analysis. PCR duplicates were also removed

and the remaining 3,767,750 reads were mapped to the transcriptome using the TopHat aligner (Trapnell et al. 2012). Fragments per kilobase per million reads (FPKM) values were then calculated with the Cufflinks package and the data was converted from the bam file format into the BigWig format using Samtools (Li et al. 2009) and uploaded into the UCSC Genome Browser for visualisation.

For the HCT116 cells, a dataset from the Gene Expersion Omnibus (GEO) database was selected (accession number GSM855450). This dataset contained 28,208,553 36-bp reads, of which only 1.17% mapped to ribosomal genes. The difference with RPE1 cells is likely due to the fact that RNA in this sample was prepared via positive selection for polyA RNA species, rather than rRNA depletion. Following PCR duplicate removal, 11,444,955 reads remained and were analysed in the same manner as the RPE1 sample, using TopHat, Cufflinks and Samtools. As more reads were analysed for the HCT116 cell line, the genome coverage for that sample was higher.

3.4.1 Correlations between transcriptional levels and fragility in RPE1 cells

I first examined the relationship between steady state transcription levels and fragility in RPE1 cells. I selected the 13 fragile cytogenetic locations I defined in the RPE1 cell line (Table 3-1) and investigated how the distribution at FPKM values within those regions related to the frequency of breaks. If expression levels of transcripts across the region contribute directly to fragility in a simple, linear manner, it would be expected that locations where most breaks occur would also have the highest median levels of expression. I did not observe this tendency in the RPE1 cell line (Figure 3-17). The most fragile location in RPE1 cells, FRA1C at 1p31.2, contained no annotated transcripts. Another highly fragile CFS, FRA2F at 2q22.2, where 8.5% of all breaks occurred, also did not contain any transcribed genes. For the remaining locations, there was no correlation between fragility and median transcriptional levels across the region (measured in FPKM) and fragility. In fact, loci with lower frequency of breaks such as FRA5C and FRA7G contained more highly

expressed transcripts than highly fragile locations such as FRA3O. To assess whether single, highly expressed transcripts contribute to fragility, I also examined whether the highest FPKM within each CFS region correlated with fragility (Figure 3-17). Again, I found no evidence to support this hypothesis – the maximum FPKM values for each region showed no correlation with fragility.

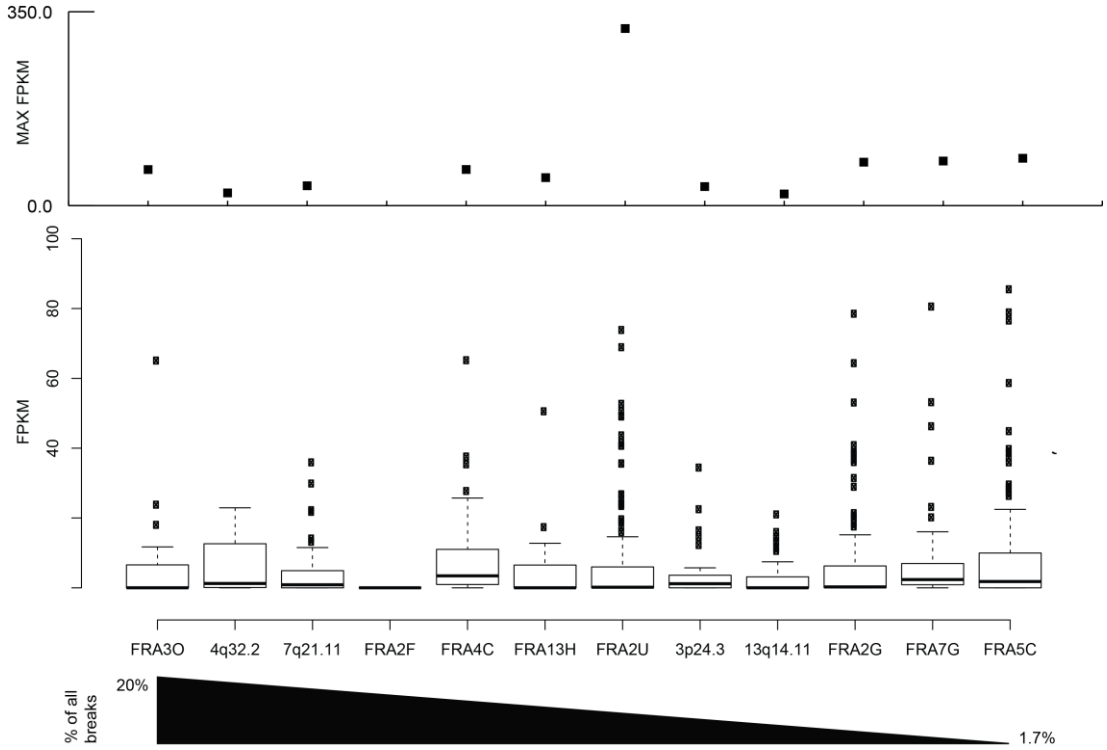


Figure 3-17 Relationship between transcription levels and fragility in the RPE1 cell line. Top panel shows the FPKM value for the most highly expressed transcript within each active fragile site in the RPE1 cell line, with CFSs ranked according to fragility levels. The bottom panel shows boxplots of all FPKM values for transcripts encompassed within each of the CFS regions.

Another possibility is that expression of single, long transcripts rather than overall transcriptional levels contribute to fragility. To test this, I selected the largest transcript for each cytogenetic location and compared the FPKM values to the fragility at each site (Table 3-4). Again, there was no evidence for an increase in FPKM value corresponding to an increase in fragility, suggesting that no linear correlation exists between transcription levels and fragility.

CFS	Rank	Transcript length	FPKM
FRA3O	2	946,318	1.11256
4q32.2	3	89,327	0
7q21	4	1,436,517	1.04611
FRA2F	5	64,691	0
FRA4C	6	437,688	1.39494
FRA13H	7	466,167	0
FRA2U	8	1,162,911	0.148107
3p24.3	9	585,587	22.464
13q14.11	10	573,451	0
FRA2G	11	203,116	15.907
FRA7G	12	899,468	12.4517
FRA5C	13	416,072	0

Table 3-4 Expression levels of long transcripts at RPE1 CFS regions.

3.4.2 Transcriptional Correlations at HCT116 CFSs

I performed a similar analysis in the HCT116 cells to determine if a correlation between fragility and expression levels could be observed in the context of a cancer cell line. I determined the maximum and median FPKM values for all transcripts spanning fragile regions from the HCT116 CFS repertoire. Similarly to the RPE1 cells, I found no evidence for correlation between transcriptional levels and fragility. FRA5C, which was ranked as the 8th most frequent site of breakage and represented just 5.2% of all breaks, contained the most highly expressed transcripts. Sites like FRA2I and FRA4F, which were the 2nd and 3rd most frequent, had very low levels of transcription. The highest FPKM value within a region also showed no correlation with fragility (Figure 3-18). When the expression levels of the longest transcripts within each region were considered, there was still no clear correlation between transcription and fragility (Table 3-5). Notably, FHIT, spanning the most fragile CFS, FRA3B, was expressed at higher level than long transcripts at the other less frequent CFSs. That tendency was not observed for other frequent CFSs in the HCT116 cell line, such as FRA4F and FRA2T, where the longest transcripts were expressed at extremely low levels.

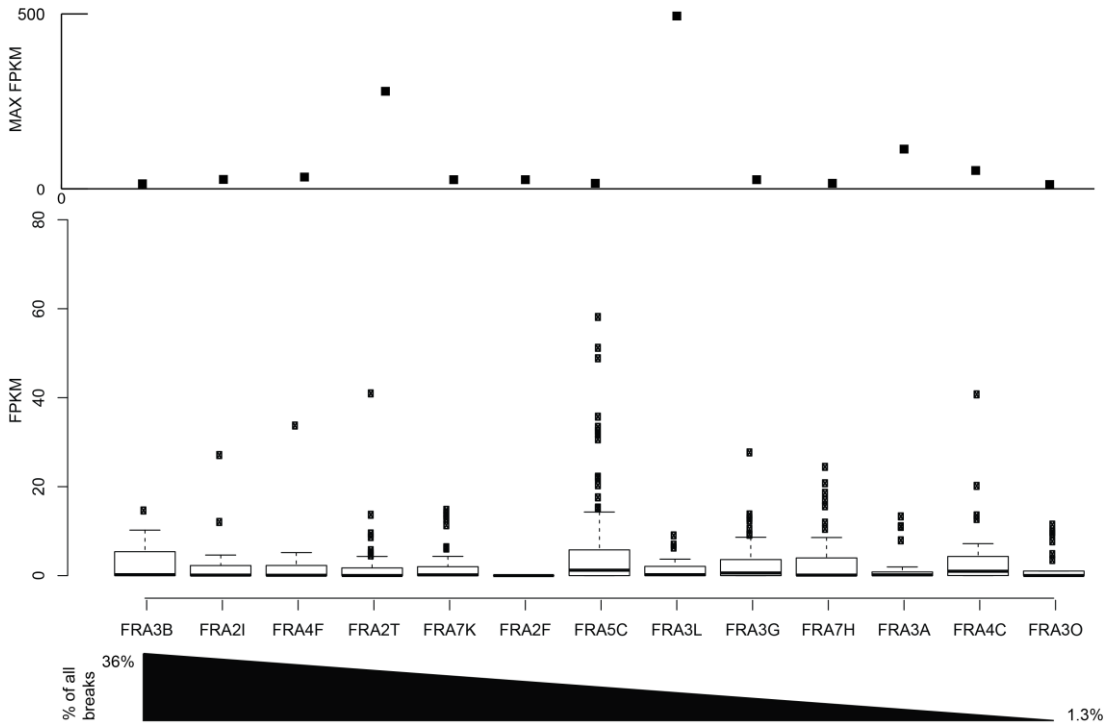


Figure 3-18 Relationship between transcription levels and fragility in the HCT116 cell line. Transcription levels across a number of fragile locations in the HCT116 cell line were investigated. . Top panel shows the FPKM value for the most highly expressed transcript within each active fragile site in the RPE1 cell line, with CFSs ranked according to fragility levels. The bottom panel shows boxplots of all FPKM values for transcripts encompassed within each of the CFS regions.

CFS	Rank	Transcrip length	FPKM
FRA3B	1	1502098	14.6316
FRA2I	2	203116	1.93857
FRA4F	3	1474687	0.165873
FRA2T	4	467341	0.00414
FRA7K	5	899468	6.37539
FRA2F	6	164691	0
FRA1C	7	NA	NA
FRA5C	8	311401	3.26769
FRA3L	9	809181	2.4204
FRA3G	10	367469	1.0942
FRA3A	11	387513	10.9447
FRA4C	12	437688	2.62547
FRA3O	13	946318	0.27698

Table 3-5 Expression levels of long transcripts at HCT116 CFS regions.

3.4.3 Transcription as a determinant of instability at CFS

The analysis above indicates that no linear correlation exists between levels of transcription at a CFS and its fragility. However, it is possible that transcription still plays a role in determining fragility at CFS loci. I exploited the differential CFS repertoires of the HCT116 and RPE1 cell line and compared the transcriptional landscapes across the most fragile locations in the two cell lines (Figure 3-19). If transcription is necessary, or plays a substantial role in inducing CFS fragility, I would expect to observe higher levels of transcription across a CFS in the cell line where the site is active. Comparing the transcriptional landscape across four common CFSs in the RPE1 cell line reveals that three of them do show higher transcriptional levels in RPE1 cells compared to HCT116 cells. This was not case for FRA2F, which displayed extremely low levels of transcription in both cell types. This trend could not be confirmed in the HCT116 cell line: three of the most fragile CFSs in that line showed lower overall levels of transcription in HCT116 than in RPE1 cells. The most fragile location in the HCT116 cells, FRA3B, showed higher levels of transcription in HCT116 than in RPE1 cells.

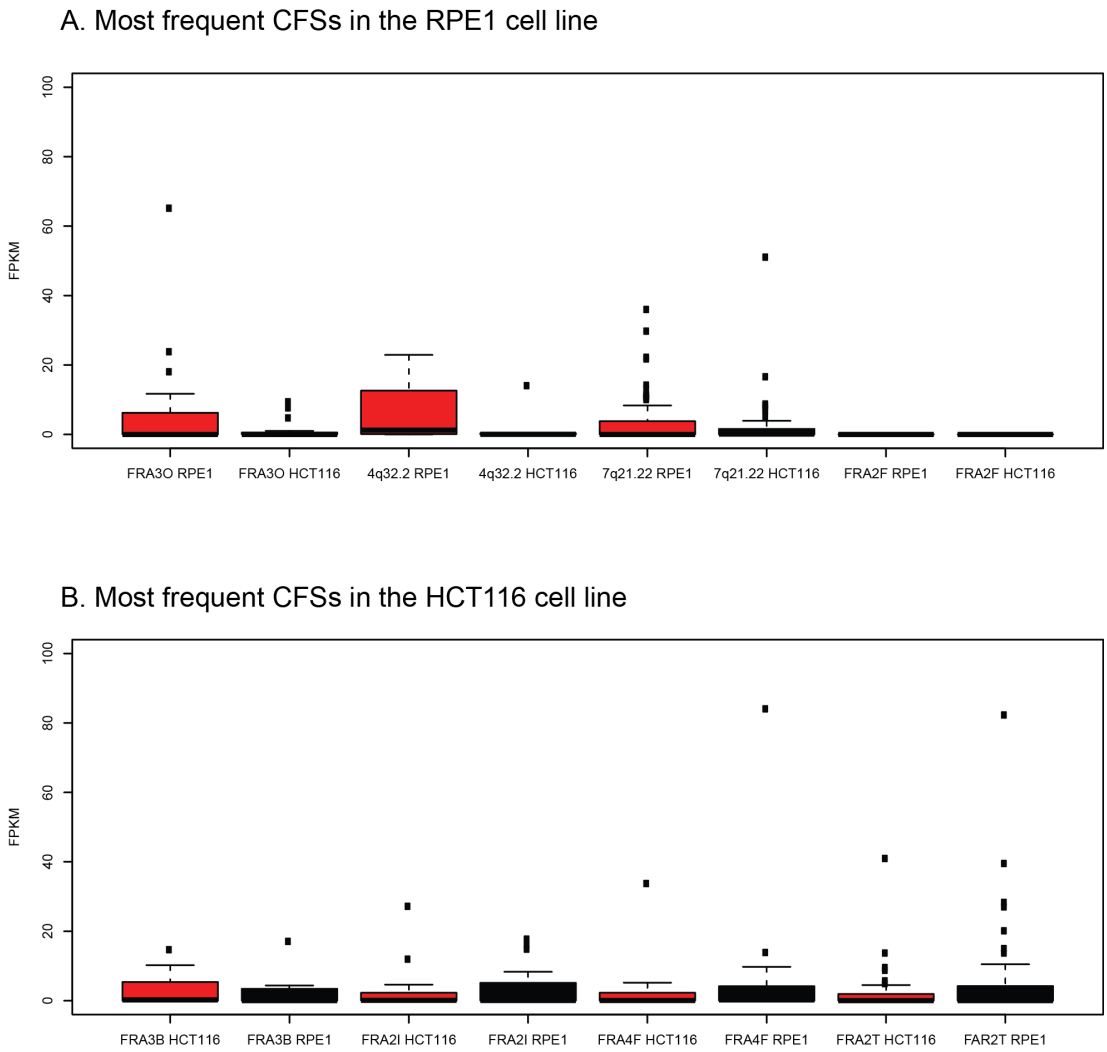


Figure 3-19 Transcriptional levels across active and inactive CFS regions. Expression levels (FPKM) for transcripts within the most frequent CFSs in RPE1 (A) and HCT116 (B) cells. Boxplots for the cell line where the site is active are shown in red and boxplots for the cell line where the site is inactive are shown in black.

Visual examination of the RNA-seq tracks at the FRA3B site confirmed that FHIT was expressed in the HCT116 cell line and not in the RPE1 cell line (Figure 3-20). While this is an interesting property of the FRA3B site, such tendency is not observed for other CFS regions in the HCT116 cell line. HCT116 CFSs with high fragility containing long genes included FRA4F and FRA3A. FRA4F contains the long genes GRID2 (1.46 Mb) and CCSER1 (1.47 Mb). No transcription was seen across the gene bodies for either of these two genes in the HCT116 and the RPE1 cell line. CCSER1 appears to show extremely low expression in a number of tissues, including testis and brain

tissues. GRID2 codes for a protein from the glutamate receptor family, which is only expressed in cerebellar Purkinje cells. FRA3A contains ZNF385D, a 0.93 Mb long gene, which was also not transcribed in either of the two cell lines. In the RPE1 cell line, higher expression levels were observed at FRA3O and the novel sites 4q32.2-4q32.2 and 7q21.12. However, this was associated with shorter transcripts, rather than a single long transcript as in the case of FRA3B. Therefore the data does not support a simple model where transcription of long genes is necessary or contributes to fragility. However, it is possible that transcription plays a role in a subset of CFS locations, such as FRA3B.

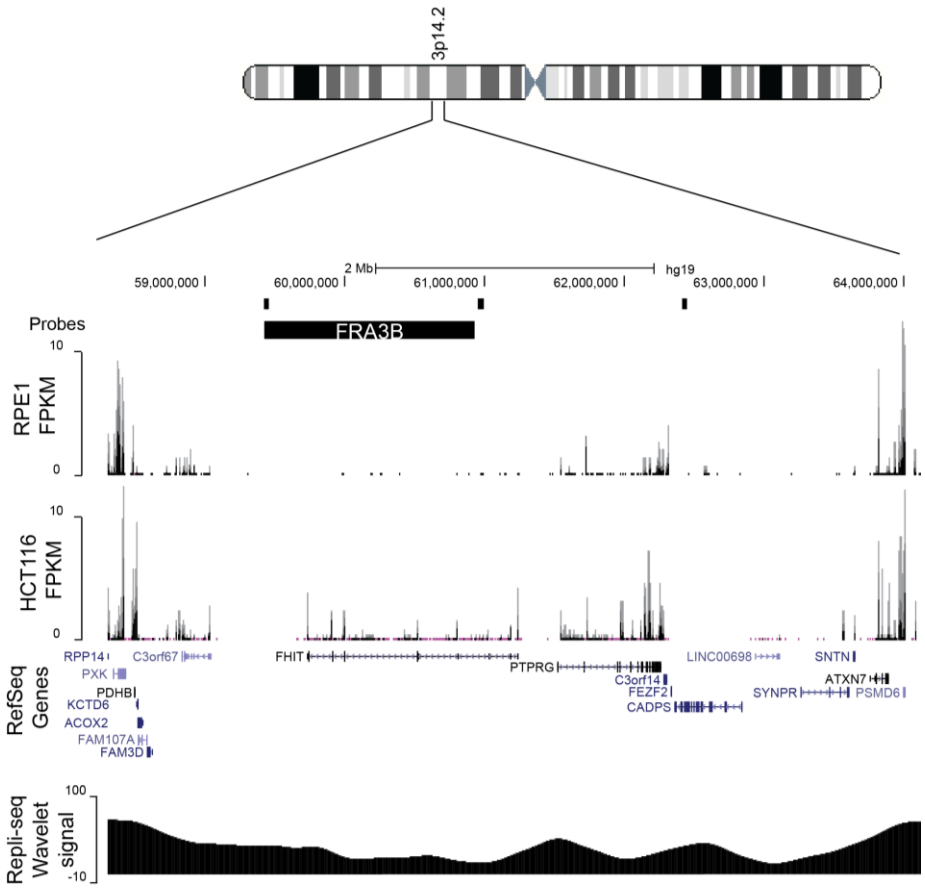


Figure 3-20 Transcriptional landscape across the FRA3B CFS in RPE1 and HCT116 cells. The positions of the fosmid probes used to fine-map the site are shown in black. The region where breaks occur is marked by a black bar. FPKM values for the RPE1 cell line (top panel) and the HCT116 cell line (2nd panel) are shown. A difference in the expression levels between the two cell lines can be seen at the long FHIT gene, which overlaps with FRA3B.

3.5 Modifying transcriptional levels at the FRA3B site

Although RNA-seq data did not support the hypothesis that transcription contributes to CFS breakage, at FRA3B, the transcriptional landscape appeared to correlate with fragility. This was an intriguing observation and I looked to explore the influence of transcription at this CFS further. I therefore used the CRISPR genome editing system to modify the expression levels of FHIT in sub-clones of the HCT116 cell line and investigated the effect of these changes on the fragility of the FRA3B locus.

The CRISPR system has been used to modify gene expression in many studies. These studies are mostly based on using CRISPR to recruit transcriptional activators (such as VP16) or repressors (such as SETB1) to defined genomic locations (Sander & Joung 2014). I avoided these approaches, as they are very transient and work through modifying the surrounding chromatin structure. Although changes in transcriptional levels are likely to be accompanied by changes in chromatin structure, I hoped to avoid large-scale alterations and preserve the local chromatin context at FRA3B, since I aimed to measure the direct effects of transcriptional levels on fragility and avoid confounding factors. With that aim, I targeted the CRISPR Cas9 system to the FHIT promoter to induce random breaks within a population of HCT116 cells, anticipating that some of the changes will result in alterations in transcriptional levels. Screening of multiple isolated clones revealed that some clones showed differential transcription at FHIT as a result of small sequence changes, which I then used in subsequent experiments.

3.5.1 CRISPR guideRNA design

Since the experimental design was based on modifying FHIT expression via induction of promoter mutations, it was important to establish whether FHIT is transcribed from a single promoter. The UCSC Genome Browser does not list any transcripts initiating from alternative promoters. I also examined publicly available

datasets containing information on nascent RNA reads, generated by GRO-Seq (GEO Database Sample GSM1124062) and found no evidence of transcripts originating from additional promoters in the HCT116 cell line. I then focused on the promoter region of FHIT. The transcription start site was identified via the RNA sequencing. Immediately upstream from the transcription start site was a region of reversal of GRO-seq read direction from the anti-sense to the sense strand, indicating the promoter position. I selected a 100 base pair sequence centred on the transcription start site, which I input into the Zhang Lab Optimised CRISPR Design tool (Ran et al. 2013). The design tool suggested five target sites on the positive strand and five target sites on the negative strand. I then performed a BLAST search with the guide RNA sequences to ensure they were not hybridising to off-target genomic locations and assessed the predicted off-target effects for each target site. Ultimately, I selected two target sites (called 2 and 8), both of which overlapped with the FHIT transcription start site, however one was located on the positive strand and the other - on the negative strand (Figure 3-21). I then designed two complementary oligos for each target site, which included the target site sequence, surrounded by the overhangs generated by the *BbsI* enzyme used to clone the RNAs into the px458 CRISPR vector (Ran et al. 2013), Table 3-6.

gRNA	Target site	Oligo 1	Oligo 2
gRNA2	GCAATTCCCAGAA GACCCCA	**CACC**GCAATTCCCAGAA GACCCCA	**AAAC**TGGGGTCTTCTGG GAATTGC
gRNA8 Reverse strand	GCCCCTACCGTGG GGTCTTC	**CACC**GCCCCTACCGTG GGGTCTTC	**AAAC**GAAGACCCCACGG TAGGGGC

Table 3-6 Target sites and oligo sequences used to target CRISPR Cas9 to the FHIT start site. Overhangs used for cloning are highlighted in bold.

The two oligos for each target site were annealed and cloned into a linearized px458 vector, which carries the wild type Cas9 protein, as well as ampicillin resistance and EGFP for selection. Successful ligation of the annealed oligoes into the vector was verified by Sanger sequencing.

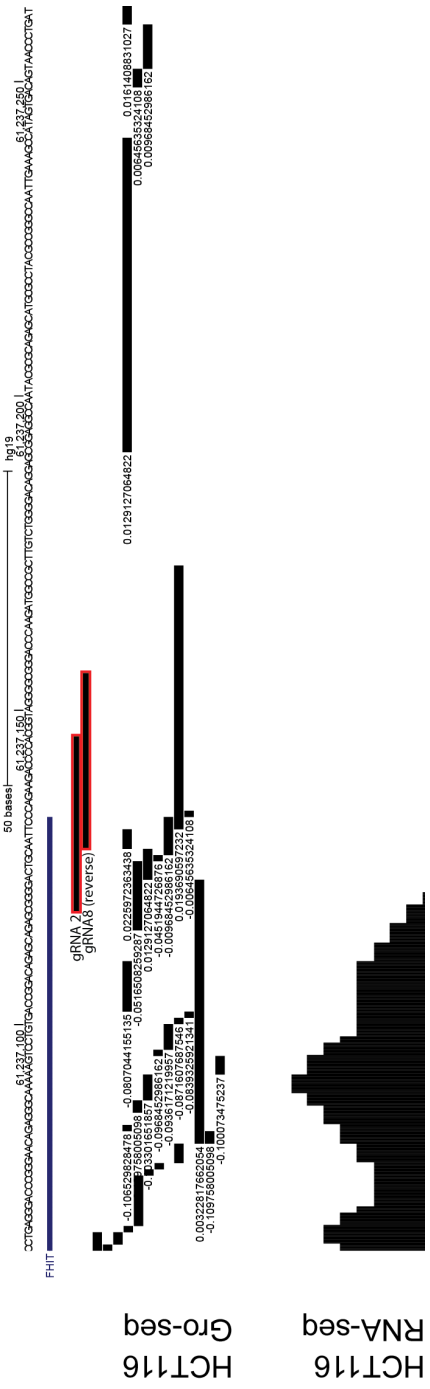


Figure 3-21 Selection of CRISPR target sites in the FHIT promoter region. HCT116 RNA-seq data is shown in the bottom panel, indicating the transcription start site. The FHIT gene track on UCSC is shown in blue. GRO-seq data for the HCT116 cell line is shown in the mid panel. Reversal of read direction can be seen just upstream of the transcription start site, indicating the promoter position. The selected guide RNAs, gRNA2 and gRNA8 are outlined in red.

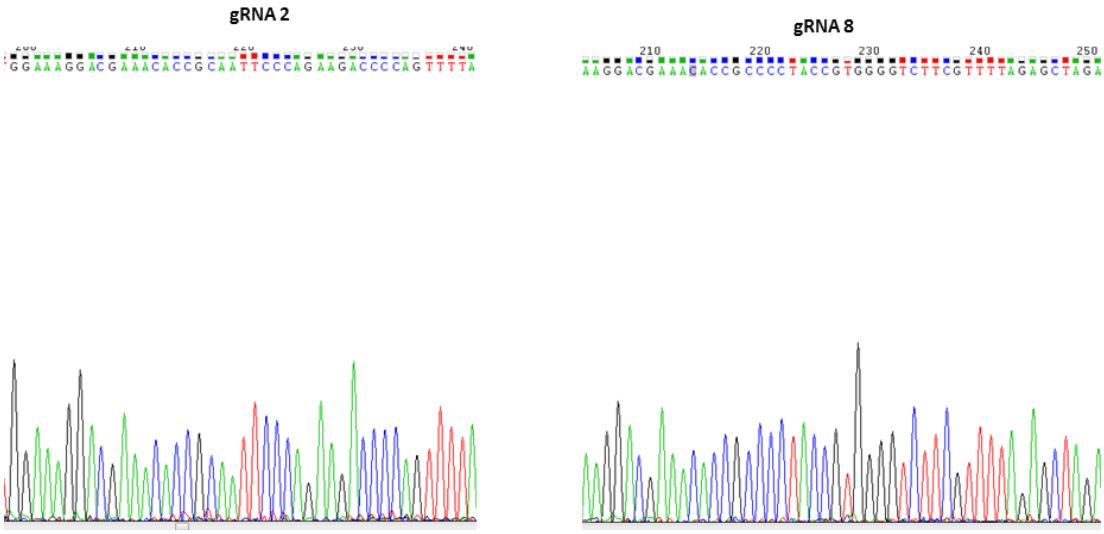


Figure 3-22 Sequencing traces showing successful ligation of the guide RNAs gRNA2 and gRNA8 in the px458 vector. Transformed DHα5 cells were plated on ampicillin plates. Colonies were selected and grown overnight and plasmid DNA was isolated. Plasmid DNA was then sequenced with the U6 forward primer (GACTATCATATGCTTACCGT), corresponding to the U6 promoter present in the px458 plasmid.

3.5.2 Assessing the efficiency of CRISPR in the HCT116 cell line

Next, I wanted to determine the efficiency of each of the two gRNAs in generating sequence changes at the target sites and examine the types of mutations that the CRISPR-Cas9 system can induce in these conditions.

To do this, I transfected the vectors containing the gRNAs 2 and 8 into HCT116 cells. Successfully transfected cells expressed EGFP and were quantified by flow cytometry, indicating that transfection efficiency was 40% for gRNA8 and 29% for gRNA2. I isolated the EGFP-positive population via FACS and grew the cells for 72 hours, allowing the CRISPR-Cas9 to induce various sequence alterations. Next, I extracted genomic DNA from this cell population and amplified a 594 bp region surrounding the target site, using primers listed in Table 3-7. If the CRISPR-Cas9 complexed with the selected gRNAs is efficient at inducing mutations at the target sites, I would expect that the amplified DNA would be made up of a pool of different amplicons with various mutations. Therefore, I performed TA cloning with the products of the PCR amplification reaction, which allowed me to isolate and

sequence individual amplicons. For the guide RNA8 samples, I was unable to find any examples of sequence changes. However, I found a range of alterations in the samples derived from cells transfected with the guide RNA2 vector. Among 78 different sequences derived from the TA-clones, 51 carried sequence changes. The most common changes were single base substitutions. Small deletions were also common and there was a recurrent 10 base pair deletion which was present in four of the sequenced samples. Larger deletions were rarer, but still present: a 49 bp and a 78 bp deletions were among the alterations in the 78 samples sequenced. Only two insertions were observed- one was 20bp long and the origin of the donor sequence could not be determined; the other insertion was 150 bp long and mapped to the U6 promoter region, indicating that a part of the px458 vector was integrated into the genome. The locations of the deletions induced by the gRNA2 targeted CRISPR Cas9 are shown on Figure 3-23. Overall, my conclusion was that oligo gRNA2 successfully guided CRISPR Cas9 to the FHIT start site, where it induced a wide range of sequence alterations with the potential to impact on transcription.

FHIT sequencing forward	FHIT sequencing reverse
GTGCGGTACAGCCTTTCGTTACACG	CTCGTGGGGCGGAAGAGTAC

Table 3-7 Primers used for sequencing the FHIT region following CRISPR-Cas9 induced mutagenesis.

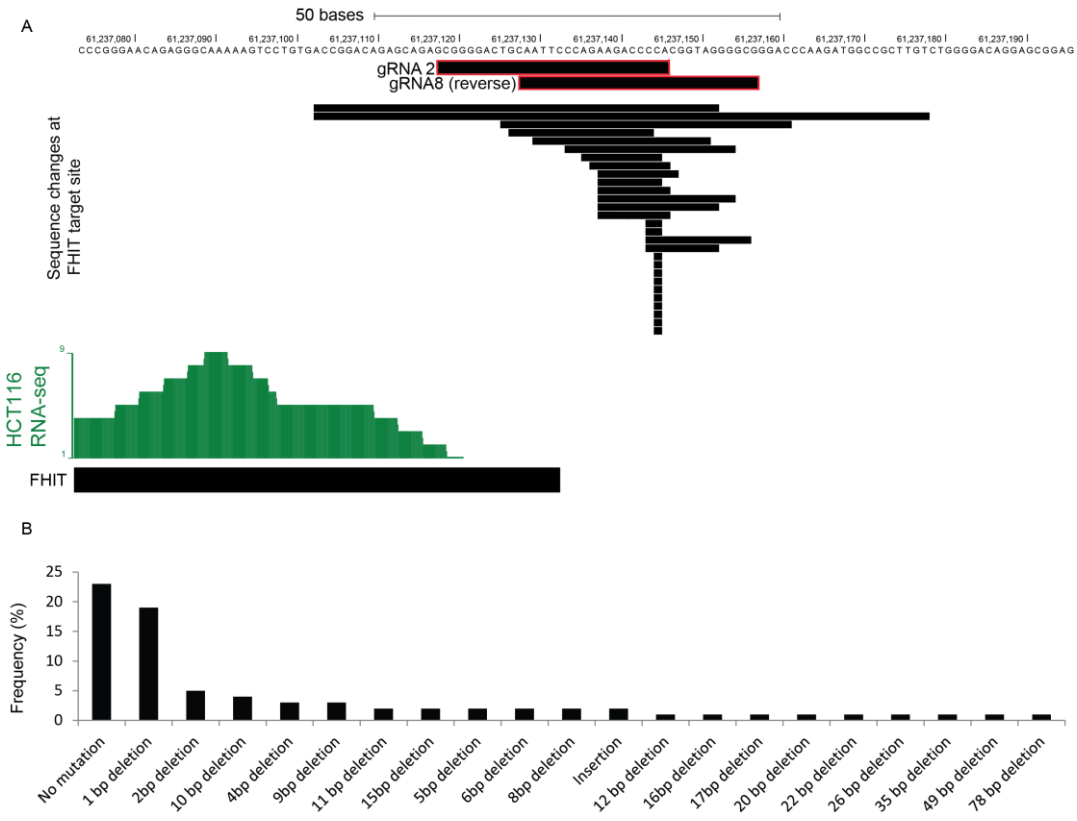


Figure 3-23 Sequence alterations at the FHIT promoter region. HCT116 cells were transfected with the px458 vector carrying the guide RNA2 or guide RNA8, which target the CRISPR Cas9 nuclease to the FHIT promoter region. Genomic DNA was isolated from the cells after 72 hours, the region surrounding the target sites was amplified and the amplicons were TA-cloned to assess individual mutations. A. Positions of the guide RNAs used in the experiment are outlined in red and shown at the top panel. The panel below shows the locations and extent of deletions induced by Cas9 targeted by gRNA2, ranging from 1 to 78 bp. RNA-seq FPKM values are shown in green below, indicating the FHIT transcription start site. B. Frequency distributions of sequence alterations in 78 TA clones derived from the HCT116 cell population targeted with gRNA2.

3.5.3 Identification of clones with differential FHIT expression

After confirming that one of the guide RNAs, gRNA2, could successfully target CRISPR Cas9 to induce a range of mutations around the target site, I grew single cell clones carrying mutations at the FHIT locus and assessed their effect on gene expression. I transfected an HCT116 cell population with the CRISPR Cas9 gRNA2 vector and allowed 48 hours for the nuclease to induce mutations. I then used FACS to isolate single cells from the transfected, GFP-positive cell population. I expanded

the single cell clones and screened them altered FHIT expression. To perform the screening, I extracted RNA from each clonal population and investigated the mRNA levels of the FHIT transcript using qPCR with three sets of primers located throughout the gene body. Screening with primers located at various points throughout the FHIT gene allowed me to determine if any alternative transcripts replaced the full-length transcript in the different clones (Figure 3-24). A total of 50 clones were screened. Clones which showed changes in FHIT expression were analysed independently at least three times. Two clones showing a consistent change in FHIT expression compared to the parental cell line were finally identified: F15 showed a reduced expression, while F3 demonstrated a consistent increase in transcriptional levels at the site.

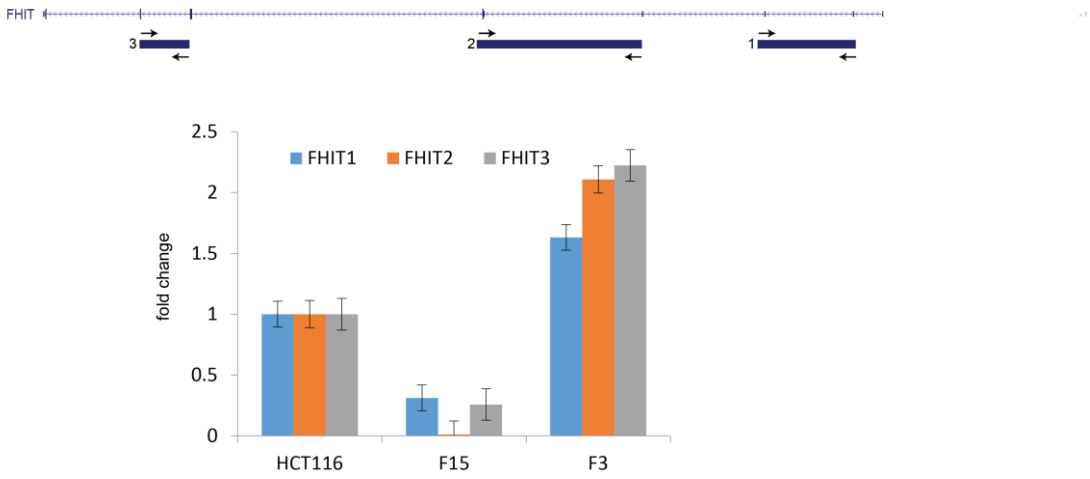


Figure 3-24 Screening clones for altered FHIT expression. Single cell clones were grown from a population transfected with gRNA2- targeted CRISPR-Cas9. The clones were expanded and screened for expression of FHIT with three different qPCR reactions spanning six exons of the gene. The positions of the qPCR primers and amplicons are shown in the top panel. The bottom panel shows fold changes for two clones showing differential expression of FHIT compared to the parental cell line.

I next examined the underlying sequence changes in the F15 and F3 clones by sequencing a 594 bp region surrounding the FHIT transcription start site using primers listed in Table 3-7. The clone over-expressing FHIT compared to the parental cells, F3, was a compound heterozygote. It carried a 10 bp deletion on one allele and a 19 bp deletion on the other. Both deletions partially overlapped with

the gRNA2 target site and were immediately upstream of the FHIT promoter (Figure 3-25). It is not clear how these deletions resulted in the consistent overexpression of FHIT. F15, the clone showing a reduced level of FHIT expression, carried a homozygous indel: 10 base pairs were deleted from the original sequence around the gRNA2 target site and a 425 bp insertion from an intronic region on chromosome 9 was present on both alleles. This is likely to have arisen from homologous recombination, with one allele first becoming mutated and then replacing the wild type allele. Interestingly, rs9880846, a SNP located 75 bp from the gRNA2 target site, was homozygous in that clone (G/G) and heterozygous in the parental HCT116 cell line (A/G), supporting the conclusion that homologous recombination took place in this clone. Again, it is not clear how the insertion of this intronic sequence resulted in a decrease of expression.

Figure 3-25 Sequence alterations in clones with modified FHIT expression. The F3 and F15 clones, showing modified expression of FHIT compared to the parental HCT116 cell line, were sequenced to determine the underlying genomic changes. F3, which shows increased FHIT expression compared to the parental cell line is a compound heterozygote carrying a 10 bp deletion on one allele and a 19 bp deletion on the other (black). F15, a clone consistently under-expressing FHIT compared to HCT116 cells, had a homozygous indel, encompassing a 10 bp deletion and an insertion of a 425 bp intronic sequence from chromosome 9. The gRNA2 target site is indicated in black and the bottom panel shows the FHIT transcription start site, as indicated by RNA-sequencing reads in the HCT116 cell line.

3.5.4 FRA3B fragility in clones with differential FHIT expression

After identifying clones with variant FHIT expression, I investigated how these changes in transcription influence the fragility of FRA3B. I treated F3 and F15 clones with 0.4 μM APH for 24 hours and prepared metaphase spreads with chromosomes derived from these cells. I determined the frequency of breaks at the FRA3B site, as well as the mean break frequency for the two samples and compared them to the frequencies in the HCT116 cell line under the same treatment conditions. I assessed 53 and 57 metaphases for the F15 and F3 clones respectively and compared them to the HCT116 data generated in the process of CFS repertoire characterisation (Section 3.1.2). In the under-expressing F15 clone, the average rate of breakage at FRA3B was 0.22 breaks/metaphase, compared to 0.19 breaks/metaphase for the HCT116 cells, while the mean number of breaks per metaphase was increased in the clonal population, at 2.86 compared to 1.88 in the parental cells (Figure 3-26). The difference in fragility at FRA3B was clearly significant (two-tailed t-test, $p = 0.02$), but contrary to what would be expected if transcription contributed to fragility, FRA3B was more fragile in this under-expressing clone, compared to the parental cells. Notably, the difference in the overall frequency of breaks was also statistically significant (two-tailed t-test $p = 0.003$). The increase in the overall fragility in these cells could be due to a number of reasons. FHIT encodes a diadenosine 5', 5’’’-P1, P3-triphosphate hydrolase involved in purine metabolism and its loss results in increased replication stress and genomic instability (Miuma et al. 2013) and it is possible that the decrease in FHIT expression could lead to an increase in fragility. Therefore, the slight increase in fragility at FRA3B in the F15 clones could reflect a general increase in CFS instability, rather than a locus-specific effect, related to the change in transcription. In F3, the clone over-expressing FHIT, the rate of breakage at FRA3B was higher than in both F3 and the HCT116 cell line (0.40 breaks per metaphase at FRA3B) and highly statistically significant (two-tailed t-test $p = 0.004$). In this cell line, the average number of breaks per metaphase was also higher than the HCT116 cells (2.5 breaks / metaphase), however that increase was not

statistically significant. Therefore, the big increase of fragility at FRA3B is likely to be locus-specific and due to modified transcriptional levels. These results suggest that there is a complex relationship between transcription and fragility at the FRA3B locus and possibly, other CFSs. The fact that fragility was preserved in the F15 cells, which express FHIT at a very low level, implies that transcription is not necessary or needed for fragility. On the other hand, the increase in the F3 population suggests that transcription can contribute to increased fragility once a CFS is active and unstable within a genomic context.

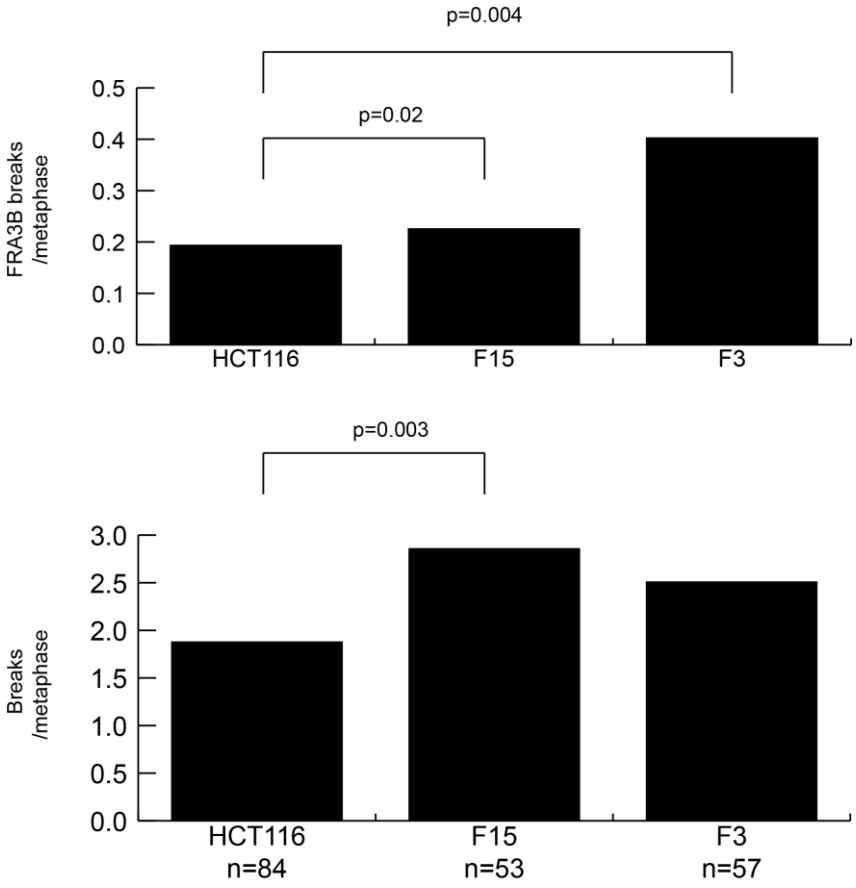


Figure 3-26 Break frequencies in clones with modified FHIT expression compared to the parental HCT116 cells. Top graph shows the break frequencies at FRA3B for HCT116 and the two modified cell lines. Both the under-expressing F15 and the over-expressing F3 showed a statistically significant increase in fragility compared to the parental cell line. p-values are given for two-tailed t-tests. Bottom graph shows the overall frequency of CFS breaks (given as breaks per metaphase) for the three cell lines. CFS fragility was significantly increased in the F15 cell line, possibly reflecting the damaging effects of FHIT under-expression.

3.5.5 Influence of transcription on mitotic chromatin structure at CFS

Finally, I wanted to investigate how a change in transcription could affect the mitotic chromatin structure at FRA3B. I hybridised probes used for fine-mapping the FRA3B region (Section 3.2.2.1) to chromosome spreads prepared from F3 and F15 cells treated with aphidicolin and quantified the frequency of atypical signals at each probe for the two clones compared them to the frequencies of atypical signals observed in the parental cell line. I assessed a total of 208 and 176 metaphases from the F3 and the F15 clone, respectively. I did not observe a difference in the frequency of atypical signals at Probe 2, which hybridises immediately upstream of exon 3 and is closest to the FHIT promoter (Figure 3-27). However, I observed an increase in the frequency of atypical signals in both clones for probe 1, which hybridises downstream from the 3' end of FHIT and flanks the fragile region in the HCT116 cell line telomerically. The increase was especially pronounced for the F3 clone, which over-expresses FHIT, where only 54% of the signals appeared normal versus 82 % in the parental cell line. Intriguingly this suggests that the increase in FHIT transcription rates results in a corresponding increase in mitotic mis-folding at FRA3B. Unfortunately, as both measurements are performed on a population level, it is impossible to determine whether this is due to an increased transcription of the gene or an increased proportion of cells expressing FHIT. Similarly to the increased rate of breaks at FRA3B in the F3 clone, this result hints at a complex relationship between transcription and CFS fragility.

In addition to mapping the frequency of mis-folding signals, these experiments enabled me to analyse the localisation of FHIT breaks in the two sub-clonal populations. In the parental HCT116 cell line, all FRA3B breaks occurred between probes 1 and 2, over the FHIT gene body. Surprisingly, there appeared to be drifting of breaks in the two sub-clones. In F3, the breaks appeared to drift slightly centromerically compared to the parental line, with a small proportion of breaks overlapping with Probe 3. The tendency for centromeric drift of breaks was even

stronger in the under-expressing F15 clone. With a single modified site, it is difficult to assess whether the break drift appeared as a consequence of the change in transcription at FHIT or due to inherent clone-to-clone variation, but again shows heterogeneity in fragile site behaviour.

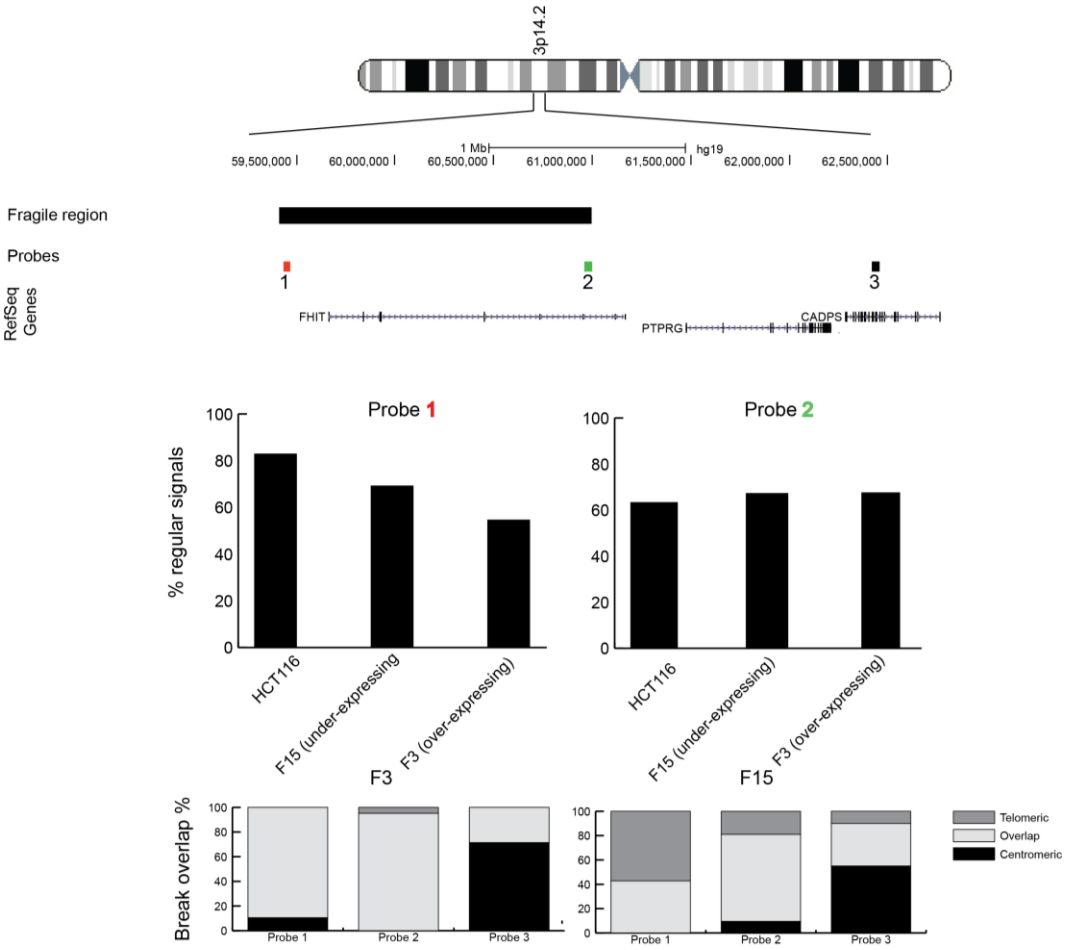


Figure 3-27 Mitotic chromatin structure in clones with modified FHIT expression. Metaphase spreads were prepared from F3 and F15 cells following treatment with 0.4 μM aphidicolin for 24 hours and hybridised to probes used for fine-mapping FRA3B. The percentage of regular signals and break overlap for each probe was quantified and compared to the parental cell line. Probe positions and the extent of the fragile region in the HCT116 cells are indicated in the top panel. The middle panel shows the frequency of regular sequence for Probe 1 and Probe 2 in the two clones and the HCT116 cell line. An increase in the frequency of atypical signals was observed for both clones at Probe 2. Bottom panel shows break positions for Probes 1, 2 and 3 in the F3 and F15 cells. The percentage of breaks where the probe is located telomerically (dark grey) centromerically (black) or overlaps (light grey) with the break is indicated for each probe.

3.6 Discussion

In this chapter, I have assessed the CFS repertoire in the RPE1 and HCT116 cell lines, the structure of mitotic chromatin at fragile regions and the contribution of transcription to instability at CFS.

I found that as expected, the RPE1 and HCT116 cell lines show differential CFS expression. Three of the fragile locations appeared to be shared between the two cell types- FRA1C, FRA2F and FRA3O, however breakage at these sites occurred at very different frequencies between the two lines. CFS mapping revealed some novel fragile sites in the RPE1 cell line, while fragile locations in the HCT116 line were consistent with previous mapping (Le Tallec et al. 2013), although at different break frequencies. Interestingly, the repertoire of fragility in both of the cell types was primarily made up of a small group of sites showing frequent, recurrent breaks combined with a larger number of locations forming breaks very rarely. Even at the most frequent sites, cytogenetic breaks were observed in only 20% of metaphases, underlying the stochastic nature of CFS breakage. Interestingly, more breaks and more severe phenotypes were observed in the tumour-derived HCT116 line, consistent with evidence that cancer cells show endogenous replication stress (Miron et al. 2015).

While many of the lesions in the two cell types presented as chromatid gaps and breaks, other types of defects, such as constrictions and concatenations were also observed. This raised the long running question of the underlying chromatin state at CFS regions: while widely accepted as breaks in the past, recent literature also speculates that these sites may represent regions of defective chromosome condensation. Decondensation effects induced by ethidium bromide and 5-azacytidine resemble fragile site lesions, while in a recent study CFSs were identified as sites of mitotic DNA synthesis and the authors speculated that mitotic compaction exposes CFS regions to facilitate repair synthesis (Minocherhomji et al. 2015). Hybridisation of BAC probes at CFS regions resulted in a high frequency of abnormal FISH signals specific to non-fragile locations, suggesting they can be used

as a tool to study defects in mitotic folding at CFS regions. The atypical signals were highly suggestive of problems such as concatenations and failure of mitotic condensation. Interestingly, the frequency of these signals on cytogenetically normal chromosomes mirrored the frequency of breaks and the domain of misfolding overlapped and extended beyond the fragile region. This suggests CFSs are located within regions prone to misfolding in mitosis, which is frequently present at the molecular level even when cytogenetic abnormalities are not observed. Presence of replication stress increases the likelihood of misfolding, although some problems with compaction persist even when cells go through the cell cycle unperturbed. I also observed the mis-compaction in both of the cell types, and across two distinct CFS loci with different features: FRA1C is gene-poor, while FRA4F spans two genes with a length over 1 Mb; this raises the exciting possibility that mitotic mis-folding is a universal and specific feature of common fragile sites. It is notable that the HCT116 cell line showed higher frequency of compaction failure, consistent with the higher frequency of breaks observed in this cell line. Triggering premature chromosome condensation indicated that at non-fragile locations, chromatin is re-set through the cell cycle to allow mitotic condensation, while this process does not happen at the FRA4F CFS locus. These observations implicate chromatin state as an important contributor to CFS fragility.

Following the determination of the cytogenetic and molecular locations for a set of CFSs, I was able to assess the contribution of transcription to fragility. Due to the significant association of CFS regions with large genes, transcription has long been suspected as a determinant of breakage at CFS. This idea is strengthened by evidence that RNaseH expression reduces CFS fragility (Helmrich et al. 2011) and that depletion of the elongation factor RECQL5 is found to induce genomic alterations at CFS regions. Confusingly, efforts to relate transcription levels to fragility have yielded negative results (Le Tallec et al. 2013). RNA-seq data from the two cell lines indicated that there was no direct relationship between transcription and fragility and that transcription was not necessary for CFS expression. However, these observations do not exclude transcription as a possible determinant of CFS

fragility. To approach the question in a novel manner, I used CRISPR-Cas9 to modify transcriptional levels at one of the best studied CFS regions, FRA3B. Genome-editing has never been used to modify and dissect the features of CFS formation to date. I was able to successfully generate HCT116 clones with modified expression of the FRA3B-associated gene FHIT and observe the effects of this alteration on break formation and mitotic folding within the region. The results were surprising. Reduced transcription did not lead to a reduction in the fragility of FRA3B, reinforcing the idea that gene expression is not required for instability. However, an increase in the expression levels led to a corresponding increase in fragility and the frequency of abnormal mitotic structures, indicating that once instability is triggered within a region, an increase in transcription can contribute.

4 Chapter 4: Characterisation of replication timing in the RPE1 and HCT116 cell line using Click-seq

Replication of the genome follows a carefully regulated temporal order, which varies between different cell types. It is not completely clear how this order is established: studies of replication timing in a range of cells of varying lineages and developmental stages show that half of the genome retains constant timing across cell types, whilst the rest of the genome shows variable replication timing across different linages (Ryba et al. 2010). It is also well known that temporal replication order is related to the functional organisation of the genome: gene-rich regions, encoding expressed genes tend to replicate earlier than heterochromatic regions. Consequently, replication timing shifts are known to correspond to changes in transcriptional state. Replication domain boundaries are also known to correspond to boundaries of topologically associated domains and are structurally established in early G1, at the same time as TADs (Dileep et al. 2015).

The careful orchestration of the replication program is achieved through regulation of origin activation. While the exact sequence and chromatin context determinants of mammalian replication origins are not defined, it is known that they frequently appear in clusters, with different origins firing stochastically at the individual cell level. Numerous origins are licensed in early G1, by the loading of the MCM2-7 complex and the assembly of the pre-replicative complexes (pre-RCs), however, only a small number of the licensed origins initiate replication in S phase, with estimates that as few as one in three to one in ten licensed origins ultimately fire (Blow et al. 2011). The licencing of a sufficient number of origins is verified at the licensing check point in G1 and if necessary, dormant origins can be recruited in response to fork stalling or DNA damage signalling (Blow et al. 2011). Firing of an origin is initiated via activation of the pre-loaded MCM2-7 complex at the onset of S-phase, which, when activated, encompasses a helicase working to unwind DNA in

front of the replication fork. Replication then proceeds bi-directionally from each origin.

The most prevalent method for studying replication dynamics is through incorporation of modified nucleotides into newly synthesised DNA. The use of modified nucleosides dates back to 1950s, when tritium-labelled thymidine was first used to track DNA synthesis via radiolabelling. The modified bases BrdU, chloro-deoxyuridine (CldU) and iodo-deoxyuridine (IdU) were later used in a similar manner, allowing detection through antibodies, which resulted in better signal and resolution. The use of multiple analogues also enabled double pulse experiments, which made studies examining fork speed and symmetry feasible. The antibody-based detection of BrdU also resulted in the development of Repli-seq, an immunoprecipitation (IP) - based method for isolating and sequencing newly replicated DNA (Hansen et al. 2010). In the Repli-seq methodology, cells are pulsed with BrdU for a short period of time and subsequently FACS sorted into different S-phase populations. Genomic DNA is extracted from the different cell stages and newly replicated, BrdU labelled DNA is enriched via IP with an anti-BrdU antibody. The isolated DNA can then be sequenced, hybridised to arrays or used in qPCR reactions to interrogate specific genomic regions. A recent improvement in the field of modified nucleotides is presented by the thymidine analogue 5-Ethynyl-2'-deoxyuridine (EdU). EdU incorporates an alkyne group which can be attached to an azide moiety in a copper-catalysed cycloaddition termed a "click" reaction (Salic & Mitchison 2008). The azide group can be linked to either a fluorescent molecule or a biotin group to facilitate either detection or isolation of newly replicated DNA. Click reactions are used widely in biology as they are highly specific, since no azide groups occur naturally in cells. EdU has therefore been widely used in imaging replication dynamics and has also been used to isolate nascent chromatin associated with newly replicated DNA (Alabert et al. 2014; Sirbu et al. 2012). However, to date, EdU has not been used to replace BrdU in the Repli-seq technique. I set out to optimise and use a variant of Repli-seq in which EdU is used instead of BrdU and have termed this technique Click-seq. In Click-seq, cells are

pulsed with EdU and FACS (fluorescence activated cell sorting) into different S-phase stages. DNA is isolated from the different cell populations and a biotinylated azide group is attached to incorporated EdU molecules in a click reaction. The biotinylation of newly replicated DNA harbouring incorporated EdU allows the antibody immuno-precipitation step from the original Repli-seq protocol to be replaced with a more robust and specific biotin enrichment. The enriched newly replicated DNA can then be analysed through qPCR or massively parallel sequencing.

I used Click-seq to analyse replication timing across the RPE1 and HCT116 cell lines, in unperturbed conditions and upon pharmacological induction of replication stress triggered by aphidicolin treatment. These experiments allowed me to investigate both the genome-wide effects of aphidicolin on replication timing as well as its locus-specific effects across active and inactive CFS regions in the two cell lines.

4.1 Optimisation of Click-seq

4.1.1 Assessment of EdU incorporation into cells

EdU was originally developed as a method for detecting nascent DNA to replace more labour-intensive and less practical alternatives such as BrdU and the radioactive analogue [3H] thymidine incorporation. As a part of the characterisaiton process, authors showed that EdU was cell permeable and was specifically incorporated into DNA during replication (Salic & Mitchison 2008). I confirmed these results in the RPE1 and HCT116 cell lines, by incubating actively dividing cells in the presence of EdU for periods from 30 minutes to one hour and then visualising incorporated EdU by clicking it to a fluorescently labelled azide. I observed distinct staining patterns, which correspond well to the patterns observed for BrdU and replication machinery components such as PCNA in cells at an early, mid or late stage of the S cell cycle phase(Dimitrova & Berezney 2002). These included diffuse staining excluded from the nucleolus and periphery, corresponding to early S-phase

cells, foci formation for mid S-phase cells and bright heterochromatic staining in late-S phase cells (Figure 4-1).

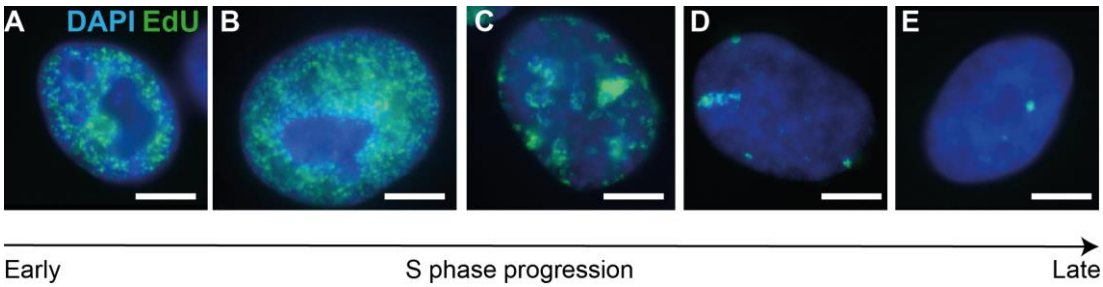


Figure 4-1 EdU staining patterns throughout S-phase. RPE1 cells were pulsed with the thymidine analogue EdU for one hour. To visualise EdU incorporation into newly replicated DNA, EdU was attached to fluorescent azide in a click chemistry reaction. The resulting patterns were indicative of the S-phase stage of the cell. A, B: Cells in early S displaying diffuse staining patterns, with signal excluded from the periphery and nucleolus. C: Mid-S phase staining with foci of active replication and signal near the nuclear periphery. D,E: Late S phase staining with signal localised to heterochromatic regions. Scale bar measuring 5 μm is shown in white.

4.1.2 EdU influence on PCR dynamics

Following the confirmation that EdU can be successfully incorporated in RPE1 and HCT116 cells, I explored how the presence of this modified nucleotide affects PCR dynamics. Specifically, I assessed if the presence of EdU in a PCR template affects its amplification rate. A substantial reduction in the amplification efficiency of EdU - containing templates would indicate that it is not a suitable nucleotide analogue to use in applications which require PCR amplification, such as massively parallel sequencing.

I first generated an EdU-containing template by amplifying a 300-bp sequence from a purified pUC18 vector, using a forward primer with a 5' GCTCACAATTCCACACAACATACGCGCC-3' sequence and a reverse primer with the sequence 5'- GCGTTATCCCCTGATTCTGTGGATAAC-3' using *Taq* polymerase and four different nucleotide mixtures. Each nucleotide mixture contained ATP, GTP and CTP in 1:1:1 molar ratios, as well as an equimolar combined concentration of EdUTP and TTP; each of the different mixtures contained varying proportion of EdUTP,

including 0%, 25%, 75% and 100% of the overall EdUTP : TTP pool. I found that no PCR products were generated for the samples where EdU exceeded 25% of the EdUTP/TTP pool. As Taq polymerase is not very efficient in incorporating modified nucleotides, I tested a number of other polymerases including *Pwo*, *KodXL* and *DeepVent exo-,* which have been shown to efficiently incorporate modified dNTPs (Tasara et al. 2003). While *Pwo* performed in a similar manner to *Taq* polymerase, I found that in reactions with *KodXL* and *DeepVent exo-* , PCR products were generated at 1:1 EdUTP to TTP ratios. I then proceeded to test whether the amplicons generated by these two polymerases contained incorporated EdU. I performed click reactions with the PCR products in the presence of biotinylated azide, which would result in the biotinylation of the amplicons. I ran the clicked amplicons on a gel and transferred them onto a nitrocellulose membrane, which I probed with a streptavidin antibody. The resulting signal confirmed that PCR products amplified by *KodXL* and *DeepVen exo-* in the presence of equimolar amounts of EdUTP and TTP contain incorporated EdU (Figure 4-2). I then proceeded to assess the impact of template EdU on PCR dynamics by using the EdU-containing PCR products as templates for subsequent amplification reactions. I performed PCR reactions from these templates using *Taq* polymerase and found that EdU in the template had no obvious impact on amplification efficiency (Figure 4-3).

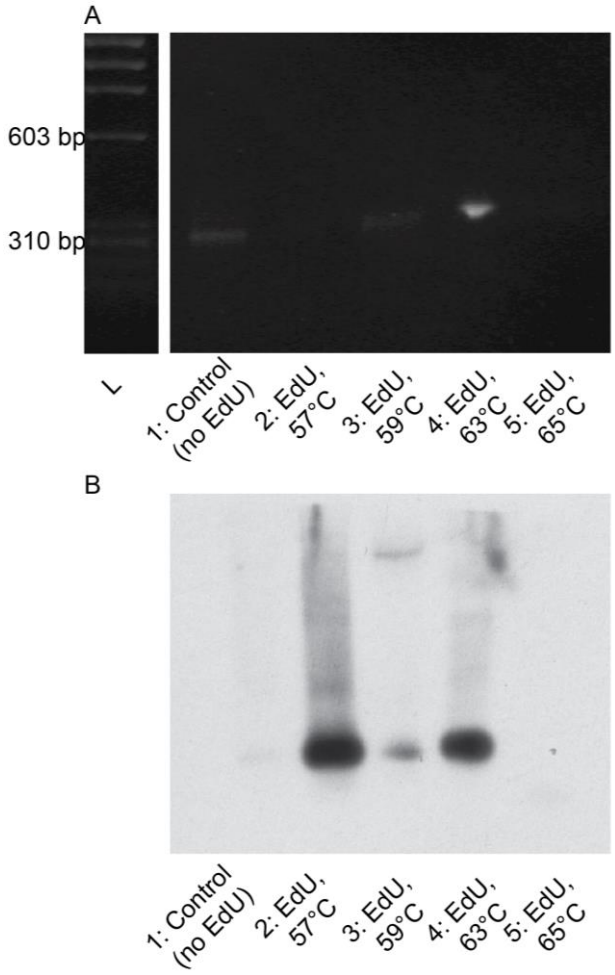


Figure 4-2 Generating PCR templates containing EdU. EdU was incorporated into PCR reactions woith the *KOD XL* enzyme, using EdUTP and TTP in equimolar ratios. The PCR products were then labelled with biotin and run on a gel. A. PCR products: L- marker, phiX DNA digested with HaeIII enxyme; 1- control PCR reaction, amplified in the presence of TTP only; 2- PCR in the presence of EdU and TTP in 1:1 ratio and annealing tempetarture of 57°C; 3- PCR in the presence of EdU and TTP in 1:1 ratio and annealing tempetarture of 59°C; 4- PCR in the presence of EdU and TTP in 1:1 ratio and annealing tempetarture of 63°C; 5- PCR in the presence of EdU and TTP in 1:1 ratio and annealing tempetarture of 65°C. B. Gel products were blotted onto a nitrocellulose membrane which was probed with streptavidin to assess the presence of biotinylated EdU. 1- control PCR reaction, amplified in the presence of TTP only; 2- PCR in the presence of EdU and TTP in 1:1 ratio and annealing tempetarture of 57°C; 3- PCR in the presence of EdU and TTP in 1:1 ratio and annealing tempetarture of 59°C; 4- PCR in the presence of EdU and TTP in 1:1 ratio and annealing tempetarture of 63°C; 5- PCR in the presence of EdU and TTP in 1:1 ratio and annealing tempetarture of 65°C. Biotin was present in all of the EdU-containing reactions and not in control reactions.

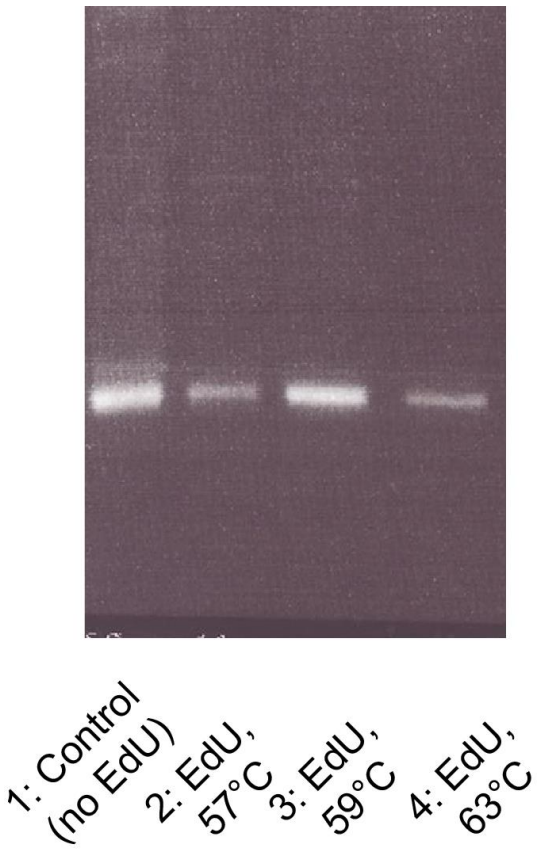


Figure 4-3 Effect of EdU-containing templates on PCR amplification. PCR reactions were set up using the EdU-containing products from the *KOD XL* reactions as template. Lane 1- PCR reaction set up with control template, which did not contain EdU; Lanes 2- 4 PCR reactions set up with EdU containing templates, corresponding to lanes 3,4 and in Figure 4-2. No difference in amplification efficiency was observed between EdU containing templates and templates generated in the presence of TTP only.

4.1.3 Implementing the EdU methodology in live cells

After I established that using EdU for labelling nascent DNA is unlikely to lead to a significant bias in the generation of sequencing libraries, I tested EdU as a replacement for BrdU in the Repli-Seq protocol in live cells. I pulsed RPE1 and HCT116 cells with EdU for one hour and 30 minutes, respectively. The different pulse times accounted for a difference in the cell cycle dynamics between the two cell lines: RPE1 cells progress through S-phase slower than HCT116 and I aimed to ensure sufficient uptake of the EdU molecule in RPE1 cells in unperturbed cells, as well as under conditions of replication stress, when less EdU is likely to be

integrated. Following the pulse, cells were harvested, fixed in 1% PFA and permeabilised as described in Section 2.7.2.

To optimise FACS (fluoresccell sorting), I determined the position of the actively replicating cell population within the flow cytometry cell cycle profile. Staining an actively growing cell population with a DNA dye such as PI results in a well-described fluorescence intensity profile when single cells are analysed by flow cytometry: a peak denotes the G1 population carrying 2n genomic content; another peak, located at approximately twice the fluorescence intensity of the G1 peak marks the G2/M population, carrying 4n karyotype; replicating cells with genomic content ranging from 2n to 4n are located between the G1 and G2 peaks. To investigate the exact location of replicating cells, I analysed an EdU-pulsed and PI-stained HCT116 sample in which I had labelled the incorporated EdU with fluorescent azide. I found that the fluorescently labelled EdU population partially overlapped with both the G1 and the G2 peak and, as expected, fully overlapped with the population located between the two peaks (Figure 4-4). However, for Click-seq, it would be impractical to fluorescently label incorporated EdU prior to cell sorting, as I found this results in a depletion of sites for subsequent biotin addition, reducing the biotinylation of nascent DNA. To encompass replicating cells located in the G1 and the G2 peak based only on PI staining and to keep gating constant among different samples, I opted for the following strategy: prior to sorting, I calculated the distance between the mid-points of the G1 and G2 peaks. I then allocated a third of the distance to each the early- S, the mid-S and the late –S sorting gate (Figure 4-4); a similar strategy has been previously employed for Repli-seq sorting (Ryba et al. 2011). The three gates reflected three incremental increases in genomic content, but I also verified that replication was equally distributed between the three gates. To do that, I quantified the cumulative EdU intensity within each gate and expressed it as a percentage of the total EdU intensity observed in the replicating population. I found that the early gate accounted for 36%, the mid gate for 28% and the late gate for 35 % of total EdU intensity,

indicating that the three gates encompassed similar percentages of total replication.

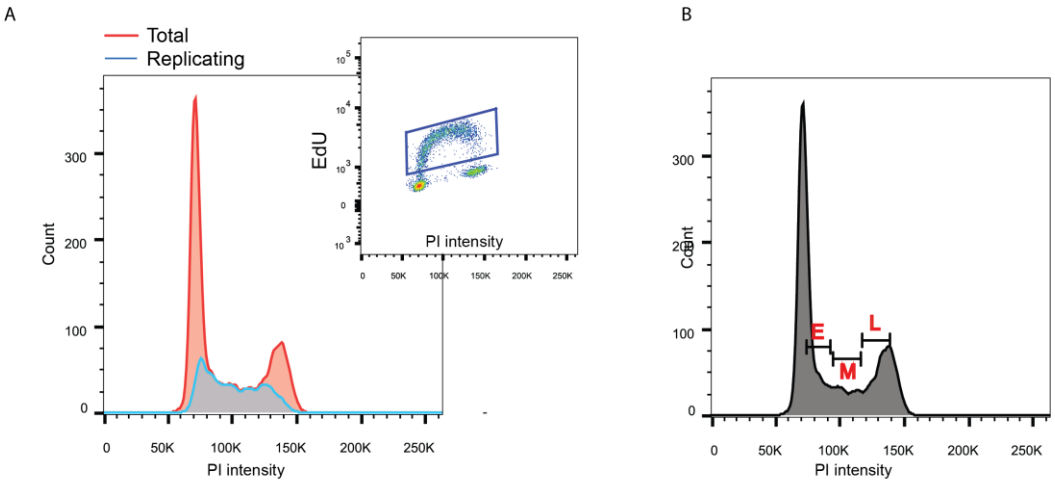


Figure 4-4 Determining gates for FACS of replicating cells. A. The localisation of replicating cells on a cell cycle profile derived from PI-stained growing cell population was determined by fluorescently labelling EdU incorporated during replication. Inset, a profile of the same population showing PI intensity (x axis) and EdU label intensity (y axis). The gate used to define the replicating, EdU-positive population is shown in blue. Overlaying the replicating population (blue) on the total cell population (red) revealed that replicated cells overlap with the G1 and G2 peaks. B. Cell sorting gating strategy for Click-Seq. To encompass replicating cells within the G1 and G2 peaks, the space between the mid-points of the G1 and G2 peaks was split into three equally sized gates, with each corresponding to early (E), mid (M) and late (L) S-phase populations.

Following FACS of an initial test sample for both RPE1 and HCT116 cells, DNA was extracted from the early, mid and late-S phase fraction for each cell type and clicked to biotinylated azide (reaction conditions described in Section 2.9.2). The addition of biotin was assessed by crosslinking the biotinylated-DNA to a nitrocellulose membrane and probing the membrane with an avidin- horseradish peroxidase conjugate. Biotinylated DNA was purified using streptavidin-coated magnetic beads. Successful enrichment was also confirmed by binding the purified DNA on a nitrocellulose membrane followed by biotin detection (Figure 4-5).

Finally, I tested whether DNA isolated from each fraction contained sequences from genomic locations known for early or late replication. I used the enriched DNA from

the three fractions for each cell line as templates in quantitative PCRs for eight primer pairs targeting known early and late replicating genes which have been previously used to perform QC for Repli-seq samples (Ryba et al. 2011). The primer pairs, the corresponding genes and the gene replication timing are listed in Table 4-1. However, as replication timing varies between cell types and has not been previously explored in HCT116 and RPE1 cells, some deviation from the assigned replication timing may be present. Primer pairs were found to produce a single band of the expected genomic size at their optimal annealing temperature.

The results from the PCR quality check indicate that the streptavidin-based enrichment of biotinylated nascent DNA is specific. DNA fragments from each of the targeted genomic regions were usually present in a single fraction or two fractions, indicating that EdU pulse combined with FACS and a biotin-enrichment for newly replicated DNA can successfully separate genomic regions according to their replication timing. For both cell types, a control sample of genomic DNA containing no EdU did not retain any biotin following the Click reaction. When the control sample was subjected to streptavidin enrichment and used as a template in PCR reactions with the QC primers sets, no products or a very small amount of PCR products were amplified.

In the HCT116 cell line, I found that results for the *HBB*, *MMP15* and *BMP1* genes were as predicted: they are classified as early replicating and the most amplified products corresponding to these regions were generated in the early sample (for *MMP15* and *BMP1*) or the early and mid sample (*HBB*). Similarly, for *DPPA2* and *SLITRK6*, classified as late-replicating regions, the most amplification was observed in the mid and late samples. *NETO1*, also a late region, showed signal predominantly in the late sample, but amplification was also present in the early and mid samples. Although *ZFP42* is noted as late-replicating region, most amplification was observed in the "early" sample. However, it is possible that the replication timing for the region containing *ZFP42* may differ in HCT116 cells. *PTGS2*

did not amplify in any of the samples for this cell line – curiously, it overlaps with a CFS region, FRA1K, which is not fragile in the HCT116 cell line.

In RPE1 cells, amplification in the *MMP15* and *BMP1* reactions was found mostly in the early samples, as expected. Surprisingly, amplification for the *HBB* PCR was found mostly in the mid sample, which may indicate differential replication timing for the *HBB* gene in this cell line. Amplification for *DPPA2*, *NETO1*, *ZFP42* and *SLITRK6* was observed mostly in the mid and late samples as expected. *PTGS2* was amplified from the mid sample in this cell line (Figure 4-5).

Gene	Gene Function	Forward	Reverse	Replication timing	Genomic location
HBB	Hemoglobin subunit beta	CCTGAGGAGAAGTCTGCCGTTA	GAACCTCTGGGTCCAAGGGTAG	early	11p15.4
MMP15	Matrix Metallopeptidase 15	CAGGCCTCTGGTCTCTGTCATT	AGAGCTGAGAAACCACCACCAG	early-mid	16q21
BMP1	Bone Morphogenetic Protein 1	GATGAAGCCTCGACCCCTAGAT	ACCCGTCAGAGACGAACTTGAG	early	8p21.3
PTGS2	Prostaglandin-Endoperoxide Synthase 2	GTTCTAGGCTGGTGTCCCATTG	CTTTCTGTACTGCGGGTGGAAC	mid-late	1q31.1 (FRA1K)
NETO1	Neuropilin And Tolloid Like 1	GGAGGTGGAATGCTAGGGACTT	GCTGAGTGTGGCCTTAAGAGGA	late	18q22.3
SLITRK6	SLIT And NTRK Like Family Member 5	GGAGAACATGCCTCCACAGTC	GTCCTGGAAGTTGAGTGGATGG	late	13q31.1
ZFP42	Zinc Finger Protein 42	CTTGTGGGGACACCCAGATAAG	AACCACCTCCAGGCAGTAGTGA	late	4q35.2
DPPA2	Developmental Pluripotency Associated 2	AGGTGGACAGCGAAGACAGAAC	GGCCATCAGCAGTGTCCTAAAC	mid-late	3q13.13

Table 4-1 QC primers for Repli-seq.

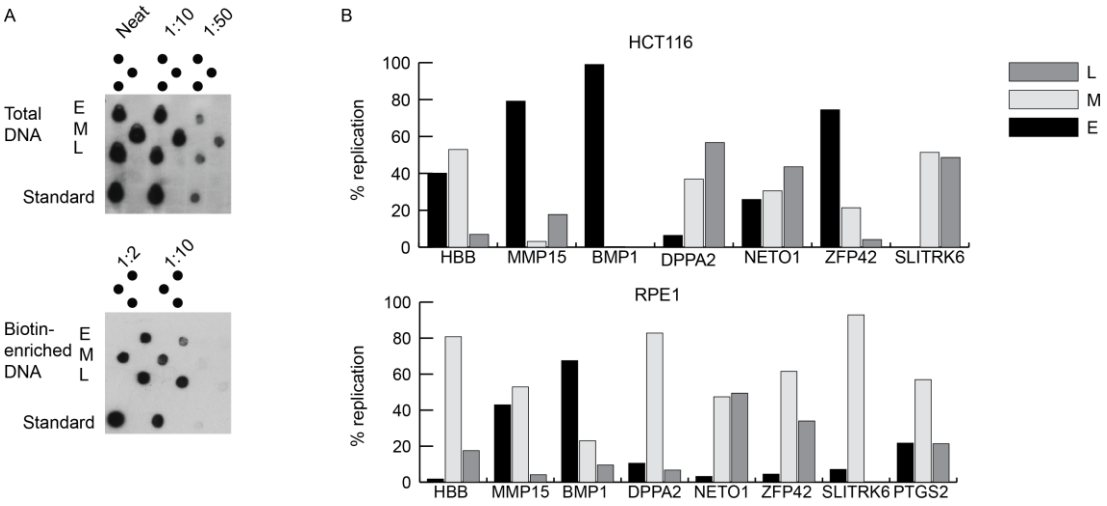


Figure 4-5 QC for EdU-based isolation of nascent DNA. A. Detection of biotin in HCT116 early (E), mid (M) and late (L) samples. Top image shows total genomic DNA after a click reaction with biotinylated azide, while the bottom image shows the same samples following enrichment of biotinylated DNA with streptavidin. Standards on both images are biotinylated T7 primers (500, 250 and 100 femtomoles from left to right). B. Results from qPCR reactions for eight genomic locations used previously in Repli-seq QC. Percentage of total replicating DNA is shown for each Early, Mid and Late samples. Percentage replication for each sample was calculated in the following manner: Fold change compared to input (pre-biotin enrichment) was calculated by subtracting the Ct value for the input sample (pre-biotin enrichment) from the Ct value of the enriched sample and using it to replace x in 2^x. For each gene, the total fold change was calculated by adding the fold change for the early, mid and late samples. Percentage replication was then calculated as the contribution of each individual sample fold change towards the total fold change.

4.1.4 Adapting Click-seq for next generation sequencing

Since results obtained with the Repli-seq QC PCR primer sets were encouraging, I optimised the methodology for use with next generation sequencing and prepared sequencing libraries from the early, mid and late fractions for the two cell lines in unperturbed conditions and following replication stress. A major point of divergence from the original Repli-seq protocol is the streptavidin enrichment replacing an antibody-based immunoprecipitation. Streptavidin-based enrichment is more robust and specific than the antibody-based IP - in fact, the original Repli-seq recommends a minimum of two hour BrdU pulse to ensure good IP effieciency. However, a minor disadvantage of the streptavidin method arises due to the high

temperature incubation in formamide required to break the biotin streptavidin bond and elute the DNA. The presence of formamide may impact on DNA quality, while the high temperature leaves DNA single stranded, which makes it inappropriate input material for Illumina NGS library preparation reagents. Some reports in the literature indicate that the biotin-streptavidin interaction can be disrupted in water, as well as formamide, at high temperatures. To determine if biotinylated DNA can be efficiently eluted from the streptavidin-magnetic beads in water, I biotinylated a sample of genomic DNA in a click chemistry reaction and performed the pull down protocol. Immediately prior to the elution step, I split the sample into two separate aliquots of equal volume and eluted one aliquot in water, while heating up to 98°C for 10 minutes, and the other - in formamide, at similar conditions. I then compared the amount of biotin present in the two eluted fractions on a nitrocellulose membrane blot. I found that equal amounts of biotin were present in the two samples, indicating that water is suitable for elution (Figure 4-6). I therefore eluted all subsequent Click-seq samples in water.

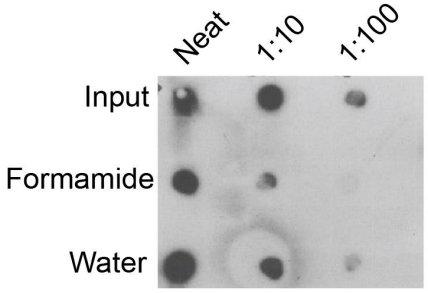


Figure 4-6 Assessing the efficiency of eluting biotinylated DNA in formamide or water. Genomic DNA containing incorporated EdU was clicked to biotin and biotinylated DNA was enriched. Prior to elution, the sample was split into aliquots of equal volume; one aliquot was eluted in water and the other-in formamide. Eluted DNA from each reaction was crosslinked on a nitrocellulose membrane, which was probed with biotin to detect DNA.

To adapt the single-stranded eluted DNA into a suitable substrate for the Illumina NGS library preparation protocol, I used a second strand synthesis reaction,

employing the mRNA Second Strand Synthesis Module from NEB, with random hexamers. The enzymes contained in that kit are DNA Polymerase I, RNase H and *E.coli* DNA Ligase. As DNA Polymerase I has no strand displacement activity, the reaction is expected to result in no amplification of the sample: instead, the polymerase syntheses the 2nd strand until it runs into a hexamer or another synthesised strand. The DNA Ligase then repairs the nicks left by the polymerase (Figure 4-7).

I processed Click-seq samples with the mRNA Second Strand Synthesis Module followed by the Ultra DNA Library prep kit by NEB, which contains components for Illumina adaptor ligation and amplification with the Illumina Universal primer, as well as the Illumina Index primers, which allow multiplexing. I found I was able to successfully generate Click-seq libraries using this protocol. I assessed the fragment distribution for the generated libraries using an Agilent bioanalyser. I found that the library fragments sizes corresponded well to the projected fragments sizes post sonication and following adapter ligation. The smooth distribution indicated there was no over-amplification of specific sequences or significant PCR biases (Figure 4-7).

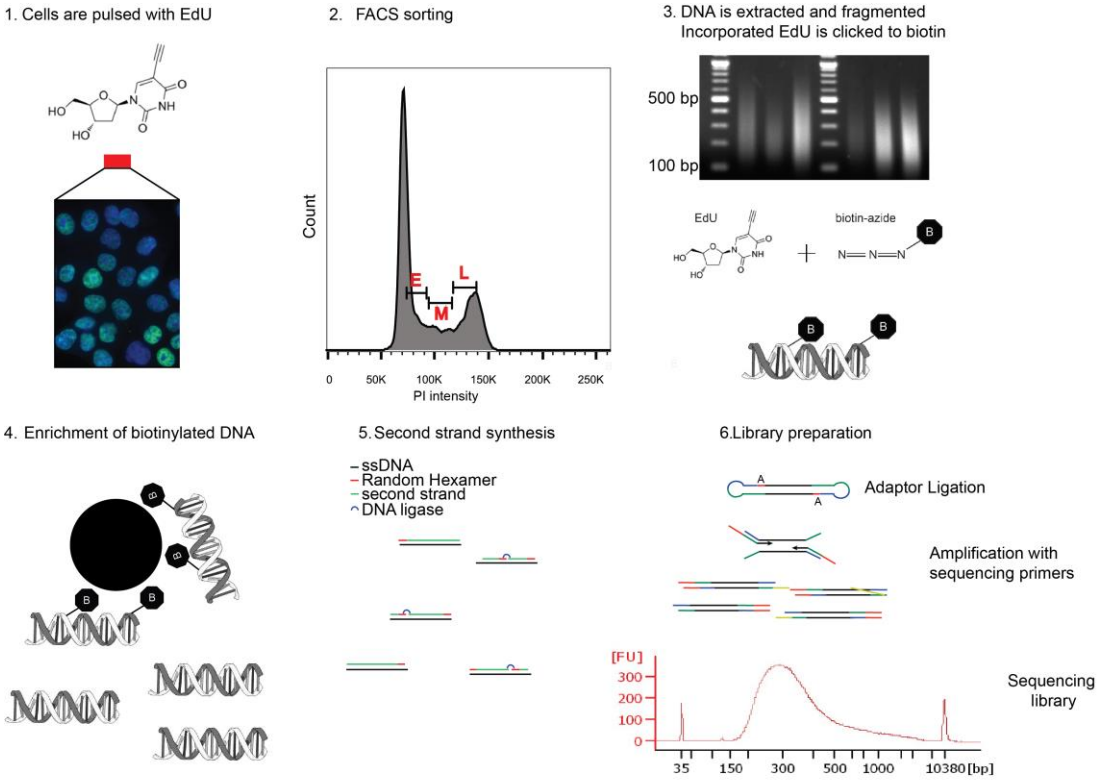


Figure 4-7 Optimised Click-seq workflow. Final version of the protocol for Click-seq sample generation for NGS. Cells are pulsed with EdU (1) and sorted into early, mid and late replication fractions (2). DNA from each cell population is extracted and fragmented to 100 – 500 bp fragment distribution (3). Next, biotin is attached to the incorporated EdU in a click reaction (3) and the biotinylated DNA is enriched by a streptavidin pull down with magnetic beads (4). The resulting single stranded DNA is used as a template in a second strand synthesis reaction (5): random hexamers (shown in red) are used to prime the synthesis of a complimentary strand by DNA Polymerase I (shown in green). As Pol I does not have strand displacement activity, synthesis stops when it runs into a DNA strand, leaving a small nick; nicks are repaired by DNA ligase (blue). Finally, double stranded DNA is ligated to the Illumina sequencing adaptor and amplified using the Illumina universal primer (red) and an appropriate Illumina Index primer (yellow) (6). The fragment size distribution of the sequencing library is assessed on an Agilent bioanalyser.

Using the optimised protocol, I prepared libraries from 24 samples from both cell lines in unperturbed conditions and under conditions of replication stress. Two biological replicates were prepared for each cell line and each condition. Details of the libraries are listed in Table 4-1. Although equal numbers of cells were harvested and prepared for cell sorting for each sample, the number of sorted cells varied substantially, from 71000 to over 2,000,000. This was due to variations in sample

quality- RPE1 cells, specifically, were found to have a tendency to form clumps during the sorting process, which reduced sorting efficiency. However, even samples with lower cell numbers were sufficient to generate high quality libraries. The 24 samples were pooled into four different pools, with six samples per pool. As per Illumina guidelines, index primers 2, 4, 5, 6, 7 and 12 were used to multiplex samples within the same pool. Prepared libraries were mixed proportionately so that the each sample had a final concentration of 5 nM per pool. The pooled libraries were sequenced on an Illumina HiSeq Sequencer by Edinburgh Genomics.

Sample	Cell type	Condition	S stage	Cell Sorted	DNA yield (μg)	Index	Sequencing Pool	Library concentration in pool
1	HCT116	Control	Early	900,000	4.4	4	HCT116_1	5nM
2	HCT116	Control	Mid	350,000	2	6	HCT116_1	5nM
3	HCT116	Control	Late	600,000	1	12	HCT116_1	5nM
4	HCT116	APH	Early	1,300,000	0.7	2	HCT116_1	5nM
5	HCT116	APH	Mid	330,000	1	5	HCT116_1	5nM
6	HCT116	APH	Late	520,000	4	7	HCT116_1	5nM
7	HCT116	Control	Early	2,500,000	7.2	4	HCT116_2	5nM
8	HCT116	Control	Mid	1,100,000	6.2	6	HCT116_2	5nM
9	HCT116	Control	Late	1,300,000	6.4	12	HCT116_2	5nM
10	HCT116	APH	Early	730,000	3	2	HCT116_2	5nM
11	HCT116	APH	Mid	1,100,000	5	5	HCT116_2	5nM
12	HCT116	APH	Late	2,200,000	13	7	HCT116_2	5nM
13	RPE1	Control	Early	1,100,000	6	4	RPE1_1	5nM
14	RPE1	Control	Mid	450,000	3	6	RPE1_1	5nM
15	RPE1	Control	Late	670,000	0.8	12	RPE1_1	5nM
16	RPE1	APH	Early	125,000	0.9	2	RPE1_1	5nM
17	RPE1	APH	Mid	71,000	0.9	5	RPE1_1	5nM
18	RPE1	APH	Late	725,000	3.2	7	RPE1_1	5nM
19	RPE1	Control	Early	1,200,000	4	4	RPE1_2	5nM
20	RPE1	Control	Mid	580,000	2.8	6	RPE1_2	5nM
21	RPE1	Control	Late	580,000	3.2	12	RPE1_2	5nM
22	RPE1	APH	Early	190,000	0.64	2	RPE1_2	5nM
23	RPE1	APH	Mid	100,000	0.7	5	RPE1_2	5nM
24	RPE1	APH	Late	450,000	2.8	7	RPE1_2	5nM

Table 4-2 Libraries prepared using the Click-seq methodology.

4.2 Sequencing results

The total number of reads obtained for each pool is shown in Table 4-3. The maximum number of reads that can be generated per flow cell on the Hi Seq instrument is approximately 200 M. Within the pools, the number of reads generated for individual samples were quite variable. Table 4-4 lists the number of reads for each of the 24 Click-seq samples: the lowest number was one of the two replicates of RPE1, mid-S fraction in the presence of APH, at 13 million; most reads were generated for a control HCT116 sample corresponding to a late-replicating fraction, at over 60 million.

Pool	Total reads
HCT116_1	129,511,895
HCT116_2	125,222,817
RPE1_1	188,463,244
RPE1_2	190,066,752

Table 4-3 Total number of reads for the four Click-seq pools

Sample	Number of reads
HCT116 E 1	19,451,240
HCT116 M 1	15,370,744
HCT116 L 1	61,241,098
HCT116 APH E 1	27,111,925
HCT116 APH M 1	35,550,942
HCT116 APH L 1	29,737,295
HCT116 E 2	24,614,757
HCT116 M 2	21,851,034
HCT116 L 2	15,909,980
HCT116 APH E 2	26,513,518
HCT116 APH M2	21,205,525
HCT116 APH L 2	15,128,003
RPE1 E 1	25,096,178
RPE1 M 1	25,005,795
RPE1 L1	23,854,537
RPE1 APH E 1	24,865,691
RPE1 APH M 1	13,292,246
RPE1 APH L 1	17,397,448
RPE1 E 2	34,603,866
RPE1 M 2	25,766,910
RPE1 L2	40,108,337
RPE1 APH E 2	32,496,647
RPE1 APH M 2	23,295,640
RPE1 APH L 2	33,795,352

Table 4-4 Number of reads obtained for each Click-seq sample.

4.2.1 Sequencing read quality

Overall quality of the sequencing runs and read quality was analysed by FastQC. The sequence quality per base was good across all samples and showed no significant deterioration throughout the run. The GC distribution over all sequences metric showed a deviation from the theoretical distribution and a sharp peak characteristic of some limited adaptor dimer contamination. Quantification of the proportion of adaptor sequences across the different samples showed that the percentage of adaptor reads did not exceed 10 % for any of the 24 samples (Figure 4-8).

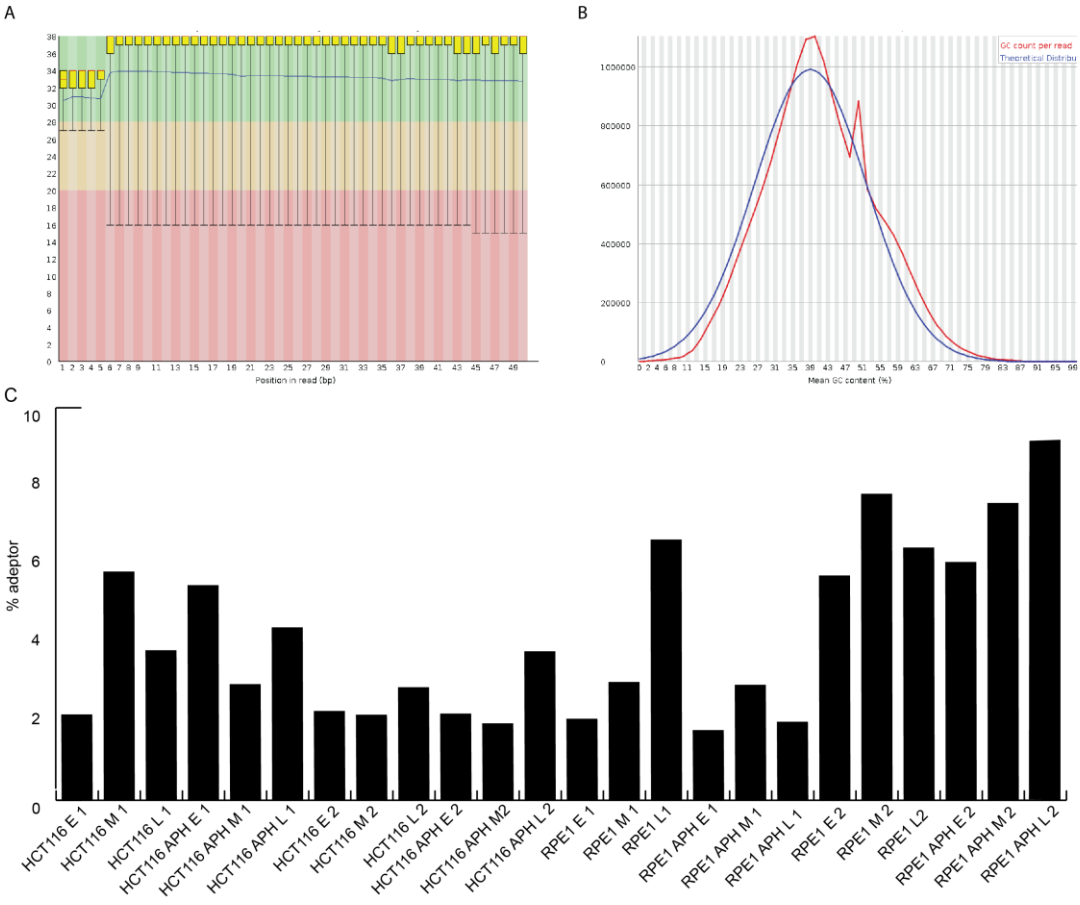


Figure 4-8 FastQC results for Click-seq libraries. A: A representative example of per base sequence quality across the reads, showing high quality was maintained throughout the run. B. An example of the GC content across reads (red) overlaid with the theoretical distribution for the human genome (blue). C. Proportion of adaptor reads across the 23 libraries. Adaptor reads did not exceed 10% of any sample.

4.2.2 Genomic alignment of reads

Following FastQC assessment, the 24 libraries were mapped to the genome with the bowtie 2 aligner(Langmead 2010). Surprisingly, a proportion of reads within each sample could not be mapped to the human genome. The proportion differed from sample to sample but was particularly high in samples treated with aphidicolin (Table 4-5). The most affected sample was one of the two duplicates for the RPE1 early fraction, in the presence of APH, for which only 22% of all reads aligned. An alternative aligner, bwa, produced similar levels of alignment (Li & Durbin 2009). In addition, alternative bowtie2 alignment options were also tested, such as "–

tryhard", which relaxes the mapping criteria. However, that resulted in a minimal increase in the percentage of mapped reads (2%). To interrogate the reads that failed to map, I extracted 100 unmapped reads for three samples with varying proportions of unmapped reads, including the most affected sample, RPE1 APH E2. I tried to manually locate these reads to all known sequences using the NCBI Nucleotide Blast tool (Altschul et al. 1990). Surprisingly, 84 of the 100 reads in that sample could not be mapped by the Blast algorithm. The remaining few reads were mostly derived from E.coli or molecular biology reagents such as vectors and phages (Figure 4-9). I also blasted a 100 unmapped reads from the sample with the highest proportion of mapped reads – HCT116 E1. In contrast to the previous sample, 53 of the 100 reads did not match any known sequence, 24 reads were from human origin and surprisingly, 23 reads came from species not directly associated with the sample preparation procedure, such as the common carp. The human derived reads in that sample carried single substitutions, were only partially aligned or were mapped to contigs removed from the hg19 assembly. Finally, I assessed a 100 unmapped reads from another aphidicolin treated sample-HCT116 APH E1, which had an alignment rate of 44%. Similarly to the RPE1 aphidicolin treated sample, a high proportion of reads failed to align to any known sequence (82). The remaining reads came from species not associated with the library preparation process, as well as a small number (5) of human-derived reads. Like in the other samples, the human-derived reads carried substitutions, mapped only partially or mapped to genomic locations which were not included in the hg19 index. A few possibilities may explain the high percentage of unmapped reads across the 24 libraries and in the aphidicolin-treated samples in particular. A simple possibility is that errors are introduced in the library preparation process: the Q5 polymerase used in the library preparation PCR reaction is an unlikely candidate as it has an extremely low error rate. DNA Polymerase I, used in the second strand synthesis step, has a higher error rate at $< 9x10^{-6}$ bases, however this would affect all samples equally. Another possibility is that aphidicolin treatment results in multiple errors during replication. In fact, a short treatment of human fibroblasts with a low aphidicolin dose resulted

in formation of deletions, duplications and novel CNVs (Arlt et al. 2009), however there was no evidence of wide-spread mis-incorporation of nucleotides. I also compared the library concentration following amplification to the proportion of unmapped reads and found there is some correlation: libraries with lower concentrations seemed to have higher representation of unmapped reads (Figure 4-9). This was especially true of the badly affected RPE1 APH E2 library, which had the lowest concentration as well as the largest representation of unmapped reads. Therefore, no conclusive explanation could be found for the higher proportion of unmappable reads in these experiments, but a low concentration and quality of starting material may be a contributor.

Sample	Total reads	Alignment (%)	Aligned reads	PCR duplicates (%)	Final number of reads
RPE1 E 1	25,096,178	78.1	19,600,115	25.40	13221655
RPE1 E 2	34,603,866	78.84	27,281,688	44.57	21,860,313
RPE1 M 1	25,005,795	75	18,754,346	67.84	5,883,239
RPE1 M 2	25,766,910	75	19,325,183	15.82	15,273,127
RPE1 L 1	23,854,537	75.55	18,022,103	12.18	15,115,571
RPE1 L 2	40,108,337	70.78	28,388,681	31.62	15,704,194
RPE1 APH E 1	24,865,691	41.02	10,199,906	23.10	4,457,218
RPE1 APH E 2	32,496,647	22.08	7,175,260	14.44	2,481,525
RPE1 APH M 1	13,292,246	51.78	6,882,725	16.88	4,638,573
RPE1 APH M 2	23,295,640	50.67	11,803,901	27.34	5,435,771
RPE1 APH L 1	23,854,537	58.8	14,026,468	0.326	9,852,726
RPE1 APH L 2	33,795,352	63.24	21,372,181	61.21	6,713,929
HCT116 E 1	19,451,240	85.73	16,675,548	17.74	13,225,572
HCT116 E 2	24,614,757	63.85	15,716,522	30.52	8,203,007
HCT116 M 1	15,370,744	68.77	10,570,461	42.08	4,103,256
HCT116 M 2	21,851,034	77.44	16,921,441	23.89	11,704,736
HCT116 L 1	61,241,098	79.44	48,649,928	29.17	30,786,004
HCT116 L 2	15,909,980	67.41	10,724,918	27.63	6,329,560
HCT116 APH E 1	27,111,925	44.35	12,024,139	22.08	6,038,060
HCT116 APH E 2	26,513,518	64.98	17,228,484	16.07	12,967,712
HCT116 APH M 1	35,550,942	58.7	20,868,403	26.03	11,613,567
HCT116 APH M 2	21,205,525	69.07	14,646,656	14.56	11,559,978
HCT116 APH L 1	29,737,295	64.09	19,058,632	20.27	13,032,514
HCT116 APH L 2	15,128,003	61.55	9,311,286	13.87	7,212,056

Table 4-5 Proportion of aligned and unique reads in the 24 Click-seq libraries.

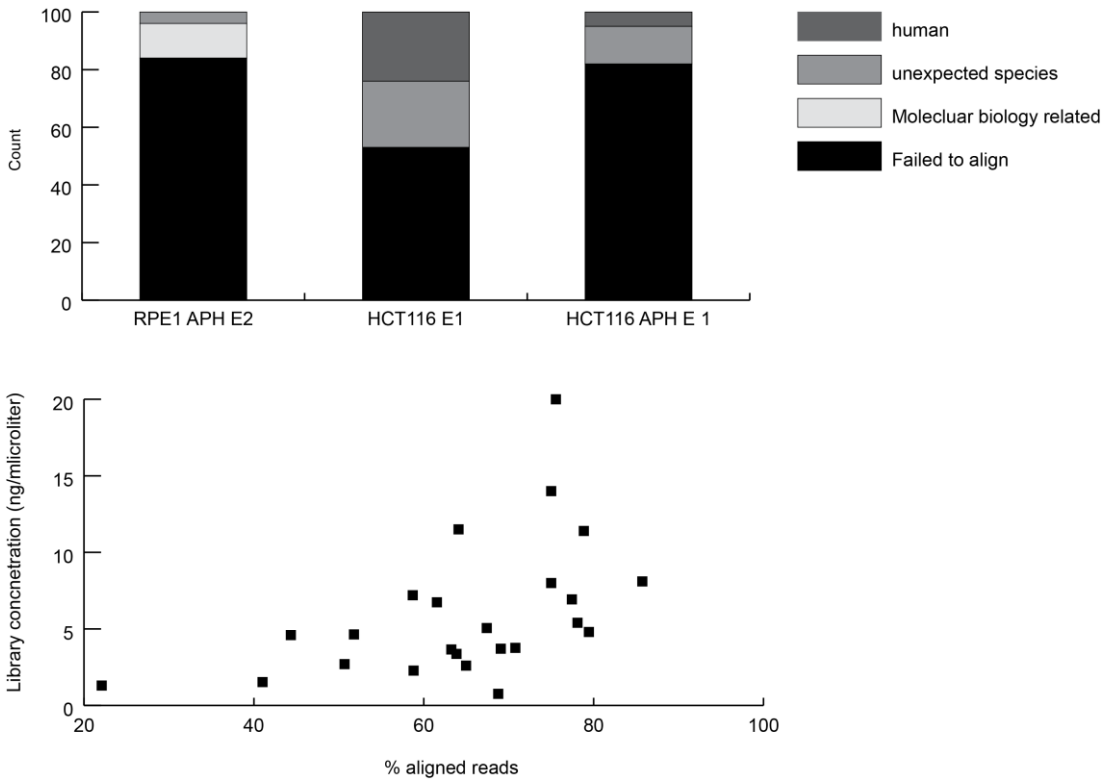


Figure 4-9 Properties of unaligned reads. Top graph: Results from blasting a 100 unmapped reads from three different samples. Most unmapped reads from all three samples failed to align to any known sequence. Bottom: a scatterplot of the library concentration (in ng/µl on the y axis) vs % of aligner reads for the 24 libraries.

Finally, following alignment, PCR duplicates were removed. The proportion of PCR duplicates also varied substantially among different samples, ranging from 12 to 67% of the reads. The final read counts for each library are shown in Table 4-5.

The final read numbers were used to calculate the genomic coverage for each sample. As each library is expected to contain only a proportion of the genome, I calculated the overall coverage for each sample, by summing the coverage achieved for the early, mid and late fraction for the sample. The final coverage for the eight samples is given in Table 4-6. It ranged from 0.24x to 0.88x per base due mainly due to the low mapping rates for the aphidicolin samples.

Sample	Resulting Coverage
RPE1 duplicate 1	0.570341
RPE1 duplicate 2	0.880627
RPE1+ APH duplicate 1	0.315809
RPE1+ APH duplicate 2	0.243854
HCT116 duplicate 1	0.801914
HCT116 duplicate 2	0.437288
HCT116+ APH duplicate 1	0.511402
HCT116+ APH duplicate 2	0.528996

Table 4-6 Genomic coverages for Click-seq samples.

4.2.3 Read counts across the genome

Finally, reads were counted in 1000 bp or 10, 000 bp windows and normalised, resulting in an FPKM value. FPKM values for the different samples were loaded as tracks on the UCSC Genome Browser and were visually assessed across the genome. Visual assessment showed that reads from each sample accumulated in visually distinct domains of sizes corresponding to known sizes for replication timing domains. At some locations, the replication wave could be followed from the early into the mid and then the late samples for each cell type. (Figure 4-10 and Figure 4-11). Obvious differences in replication timing could also be observed between the two cell types and in the presence of aphidicolin.

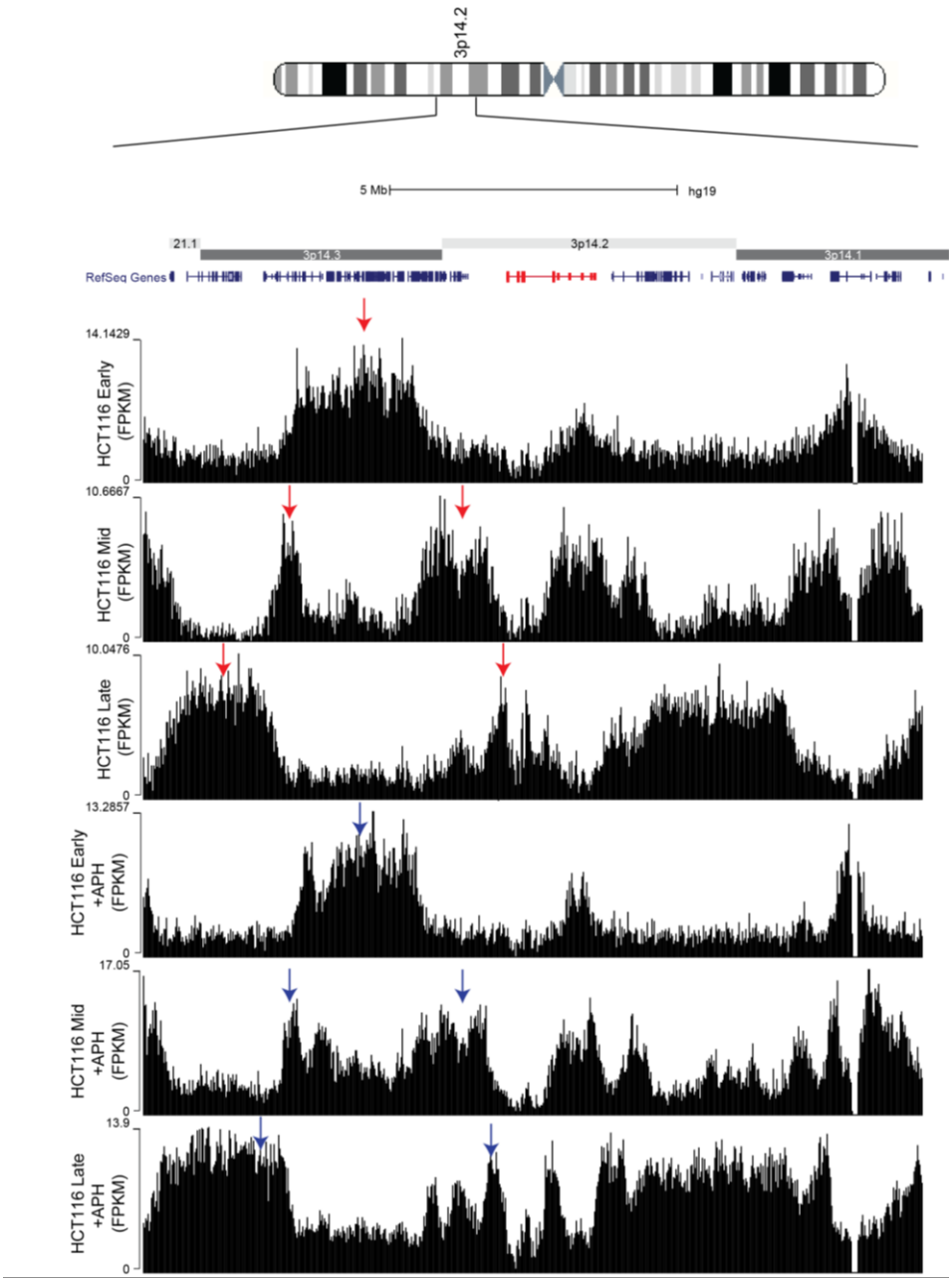


Figure 4-10 FPKM counts for the HCT116 cell line in the 3p14.1-3p14.3 region. FPKM values were calculated for all HCT116 samples and plotted across the genome. A 10 Mb region surrounding the FRA3B locus is shown here and the FHIT gene is shown in red. The top three panels represent the early, mid and late fractions for the HCT116 cell lines in unperturbed conditions, while the bottom three represent the early, mid and late fraction in the presence of aphidicolin. Arrows denote the replication wave, which can be clearly traced from an initiation zone in the early samples to putative termination sites in the late samples for both control and aphidicolin-treated samples. Small but clear replication timing differences can be seen between the control and aphidicolin treated samples.

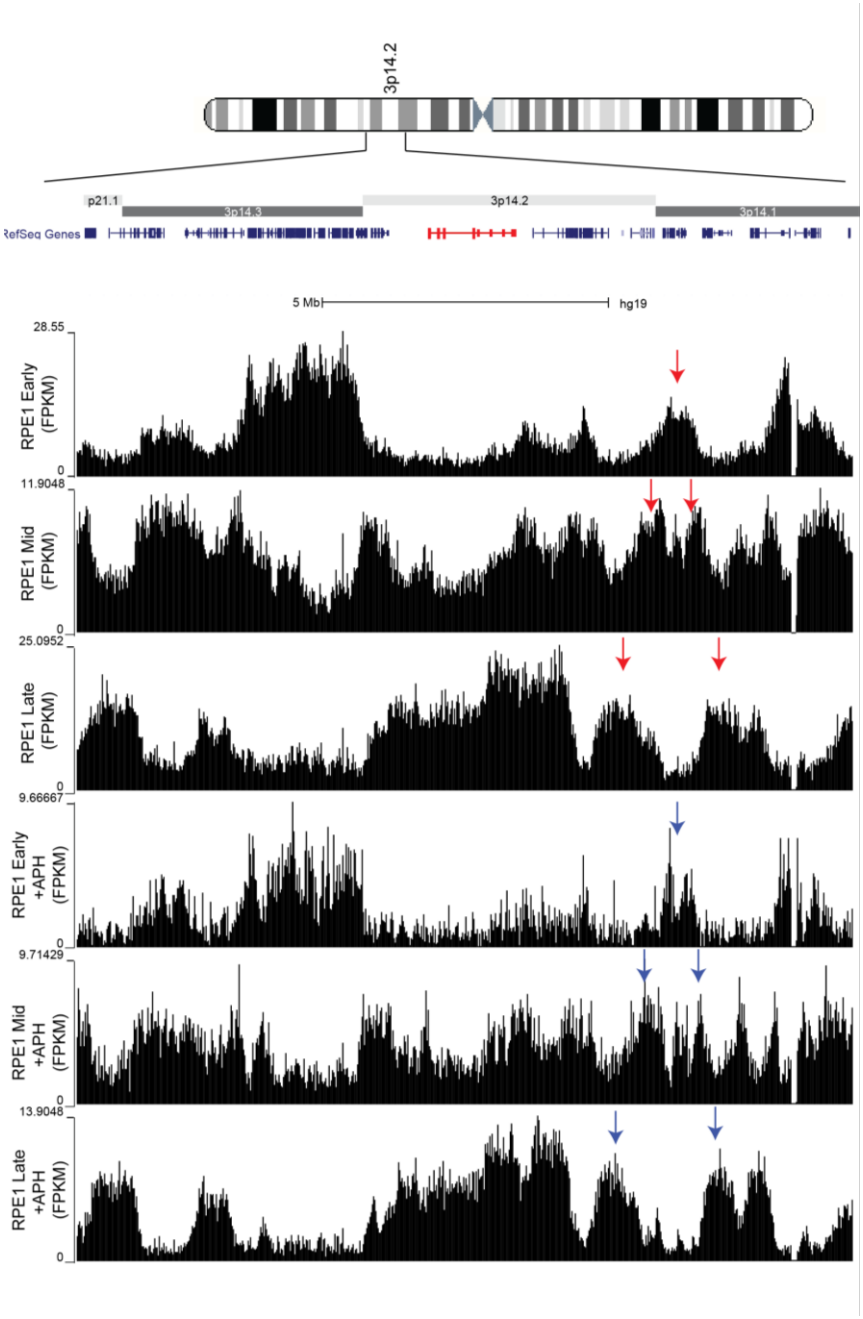


Figure 4-11 FPKM counts for the RPE1 cell line in the 3p14.1-3p14.3 region. FPKM values were calculated for all RPE1 samples and plotted across the genome. Same 10 Mb region surrounding the FRA3B locus is shown here as in Figure 4-10, with the FHIT gene location shown in red. Top three panels represent the early, mid and late fractions for the RPE1 cell line in unperturbed conditions, while the bottom three represent the early, mid and late fraction in the presence of aphidicolin. Clear differences can be seen between the replication timing of this region in the RPE1 and HCT116 cell line. The replication wave from early to late samples can also be followed in this cell type and is denoted with arrows. Like in the HCT116 cell line, aphidicolin can be seen to cause clear changes in replication timing in the RPE1 cell line.

Assessment of the read density across larger genomic regions showed an even clearer distinction of different domains within the early, mid and late samples. Visually, segments of the genome with high density of early reads overlapped with locations of high gene density, while gene-poor locations were enriched for late reads, consistent with previous observations on replication timing (Rhind & Gilbert 2013) (Figure 4-12). At this scale too, the domains across the different S-phase fractions could be seen to complement each other – domains with predominantly early reads were depleted from the late libraries, while the converse was true for late reads.

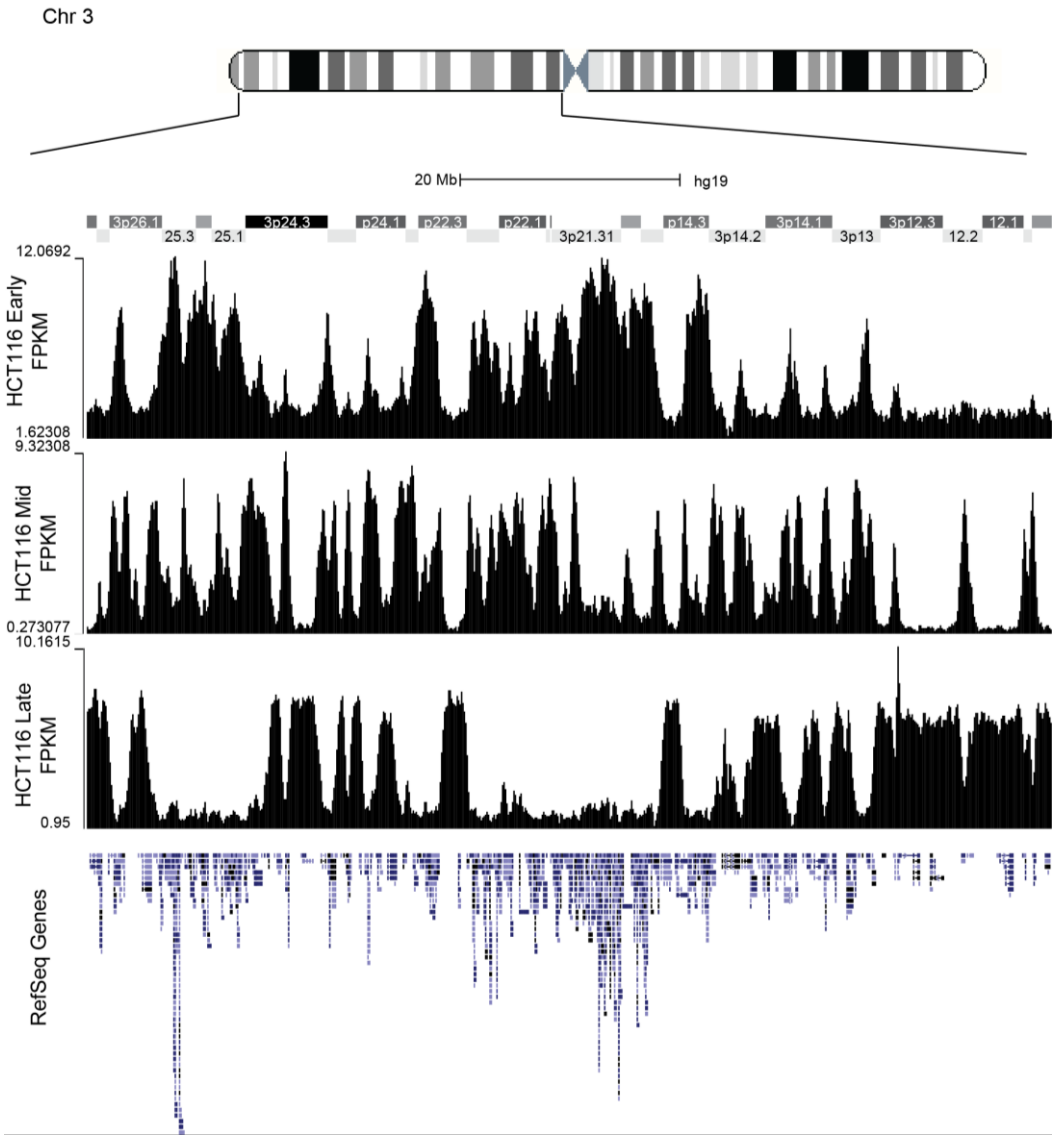


Figure 4-12 Replication timing across chromosome 3p in the HCT116 cell line. FPKM values are shown for this large genomic region in the HCT116 early, mid and late samples. High density of early reads and a depletion of late reads is found in gene-rich regions. Conversely, gene-poor regions are enriched for late reads.

4.2.4 Click-seq reproducibility across biological replicates

Next, I wanted to assess the reproducibility of the Click-seq technique across biological replicates. Visual inspection of read densities across genomic locations showed high similarity across most duplicates, with the exception of two samples from the HCT116 libraries and a single sample from the RPE1 libraries (Figure 4-13). HCT116 E rep1, HCT116 L rep1 and RPE1 M rep1 all appeared to show “flatter” read

density profiles compared to other samples and no obvious delineation of domains. However, the striking similarity between biological replicates for other samples suggests that Click-seq is a highly reproducible technique.

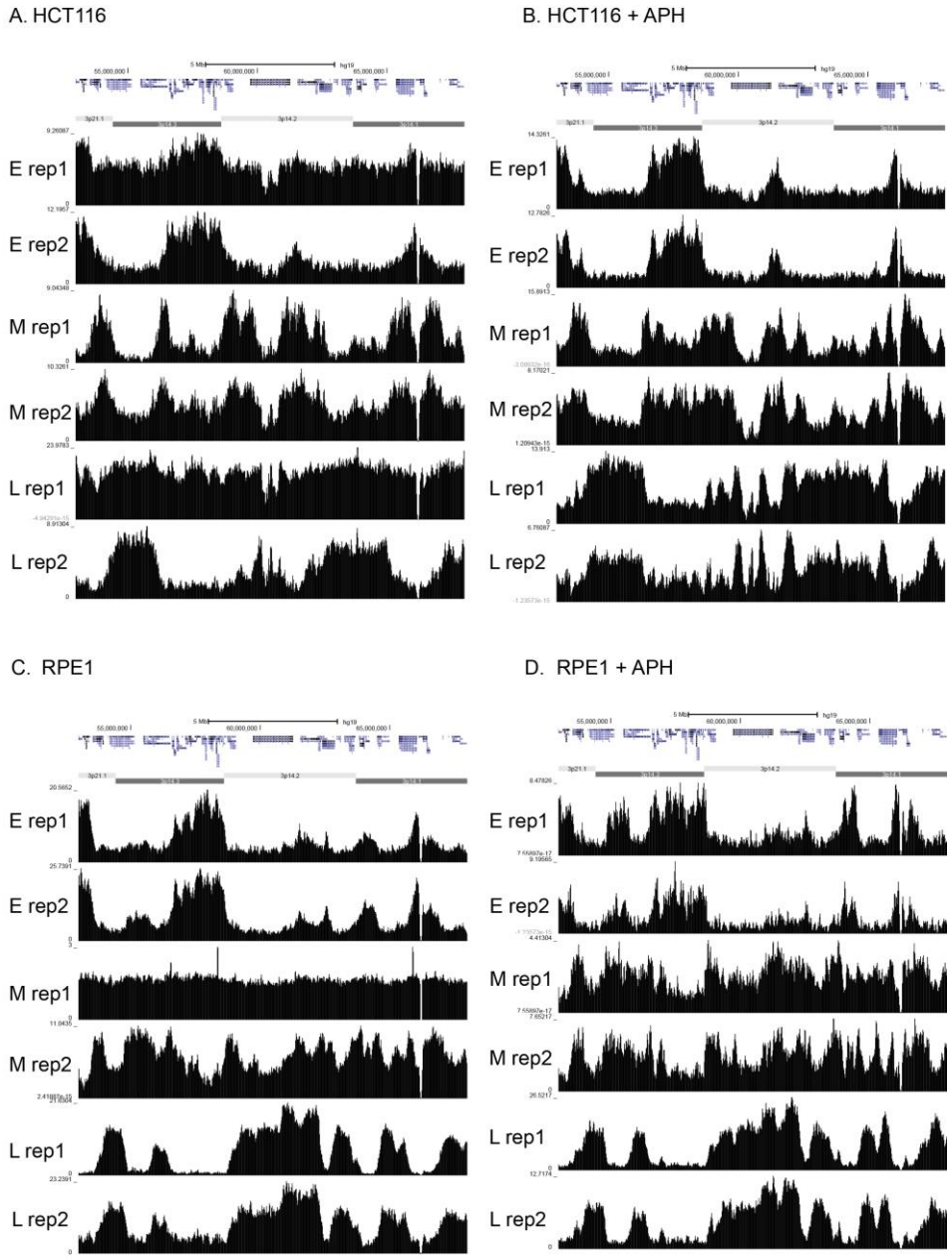


Figure 4-13 Visual comparison of read densities across biological replicates in a 15 Mb region on chromosome 3p, spanning the region across 3p14.1-3p14.3. The same region is shown with the corresponding reads for the six HCT116 libraries (A), the HCT116 + APH libraries (B), RPE1 libraries (C) and RPE1+APH libraries (D). Similarities can be observed across most biological replicates, with the exception of HCT116 E rep1, HCT116 L rep 1 and RPE1 M rep1.

I next quantified correlations between biological replicates using the R-based corrplot package. Corrplot calculates a matrix of correlation coefficient between a number of variables and visualises the correlations. Prior to using the corrplot function, I calculated the number of reads for each library in either 1000 bp or 10 000 bp windows and normalised them to the total number in the library. I then performed the correlation analysis on both 1000 bp window and 10,000 bp window datasets. The resulting correlations are shown in Figure 4-14 and Table 4-7. Overall, replicates showed very high correlations when 10,000bp windows were considered and reasonably high correlations when data was partitioned in 1000 bp windows. The increase in correlation in the 10,000 bp windows dataset may be due to a lower coverage of data in some samples, leading to more stochastic observations in the 1000 bp dataset. The lowest observed correlation between replicates was 0.5, for RPE1 APH E rep 1 and rep 2, in 1000 bp windows. However, this was increased to 0.87 in the 10,000 bp window dataset, indicating that the low correlation in the 1kb window dataset may be due to the low number of reads in the RPE1 APH E rep2 sample, which had the highest number of non-aligned reads.

Replicate 1	Replicate 2	1000 bp correlation	10,000 bp correlation
HCT116 E rep1	HCT116 E rep2	0.66	0.83
HCT116 M rep 1	HCT116 M rep 2	0.82	0.93
HCT116 L rep 1	HCT116 L rep 2	0.91	0.96
HCT116 APH E rep1	HCT116 APH E rep2	0.63	0.88
HCT116 APH M rep 1	HCT116 APH M rep 2	0.96	0.98
HCT116 APH L rep 1	HCT116 APH Lrep 2	0.9	0.96
RPE1 E rep1	RPE1 E rep2	0.79	0.94
RPE1 M rep 1	RPE1 M rep 2	0.91	0.96
RPE1 rep 1	RPE1 L rep 2	0.86	0.87
RPE1 APH E rep1	RPE1 APH E rep2	0.5	0.82
RPE1 APH M rep 1	RPE1 APH M rep 2	0.83	0.96
RPE1 APH L rep 1	RPE1 APH L rep 2	0.65	0.88

Table 4-7 Correlation coefficients between Click-seq biological replicates in 1000bp and 10,000 bp windows.

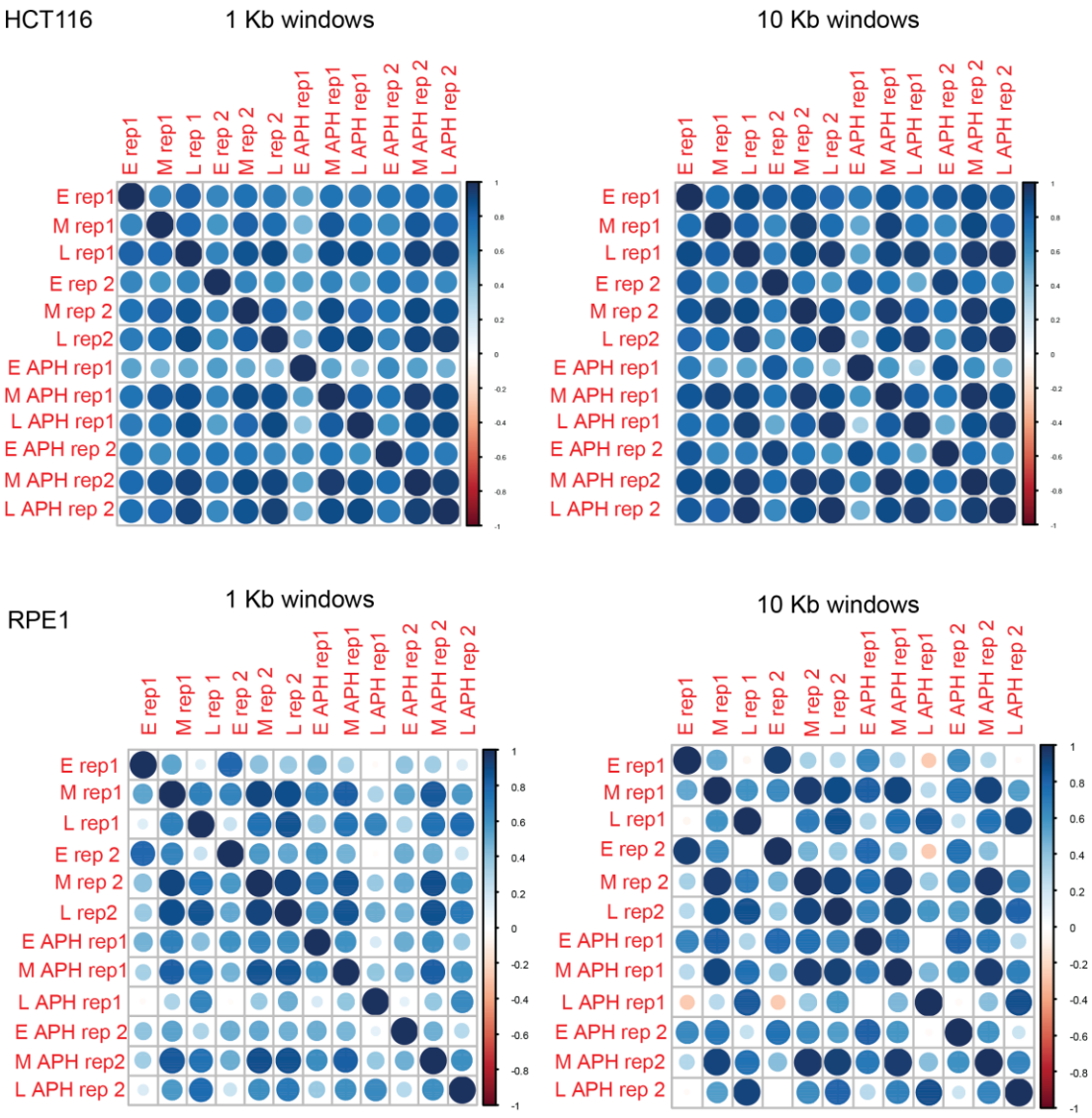


Figure 4-14 Correlation analysis for Click-seq libraries in 1000 bp and 10,000 bp windows. Plots were produced using the corrplot package in R. Darker blue indicates stronger positive correlations and the size of the circle is proportional to the correlation coefficient for two samples.

In addition to measuring the correlations for single libraries, I also combined the normalised read fragment per window values for the early, mid and late fraction for each sample into a single replication timing value (Rvalue) in such a way that windows with predominantly early reads would tend towards values of 0.8 and windows with predominantly late reads would have values close to 0.2 (described in 2.9.8). As this approach is based on assessing the proportion of early, mid and late reads from all the reads within a window, it normalises for sequence-related bias across the genome, such as GC-content and differential mappability. I assessed the

correlation of Rvalues across biological replicates for a subset of chromosomes, including chromosomes 3 and 11 (Figure 4-15). I observed good correlation between Rvalues for the RPE1 samples and the HCT116+APH samples. An exception was provided by the control HCT116 sample, where the failure of two of the HCT116 rep1 libraries could be seen to affect the Rvalues. As a result, I excluded the HCT116 rep 1 from most subsequent analyses. Overall, the high correlations observed between biological replicates indicated that Click-seq is a robust and reproducible technique.

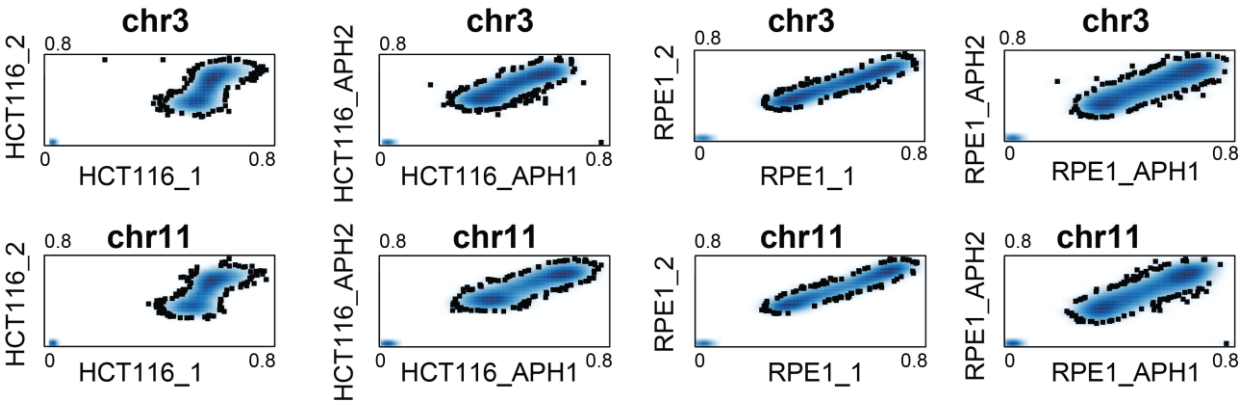


Figure 4-15 Correlations between Rvalues for biological replicates. Rvalues were calculated for each sample according to the method described in 2.9.8, in 10,000 bp windows. R values in windows were plotted for the two replicates of each sample for chromosomes 3 (top) and 11 (bottom). Good correlations could be observed for most replicates, with the exception of HCT116 rep 1.

4.2.5 GC content across Click-seq libraries

I also examined whether GC content varied across the different Click-seq libraries. Some variation between the libraries is expected, as GC-rich regions, which are associated with gene-rich portions of the genome, are known to replicate early. I calculated the number of reads across 50 bins of increasing GC content using the "read_GC" function of the RSeQC package (Benjamini & Speed 2012). This analysis was performed on the libraries following the removal of unmappable reads and included only reads that had been successfully mapped to the genome. I found some variation in GC content- as expected libraries derived from the early fractions

appeared to have higher GC content than libraries from the late fractions. An additional peak could be observed in the early samples, at around 56 % GC content. The genomic origin of reads in this peak is unclear, but it was more prominent in the untreated early samples, suggesting it may be due to a real difference in replication timing for a subset of genomic locations, rather than a technical artefact.

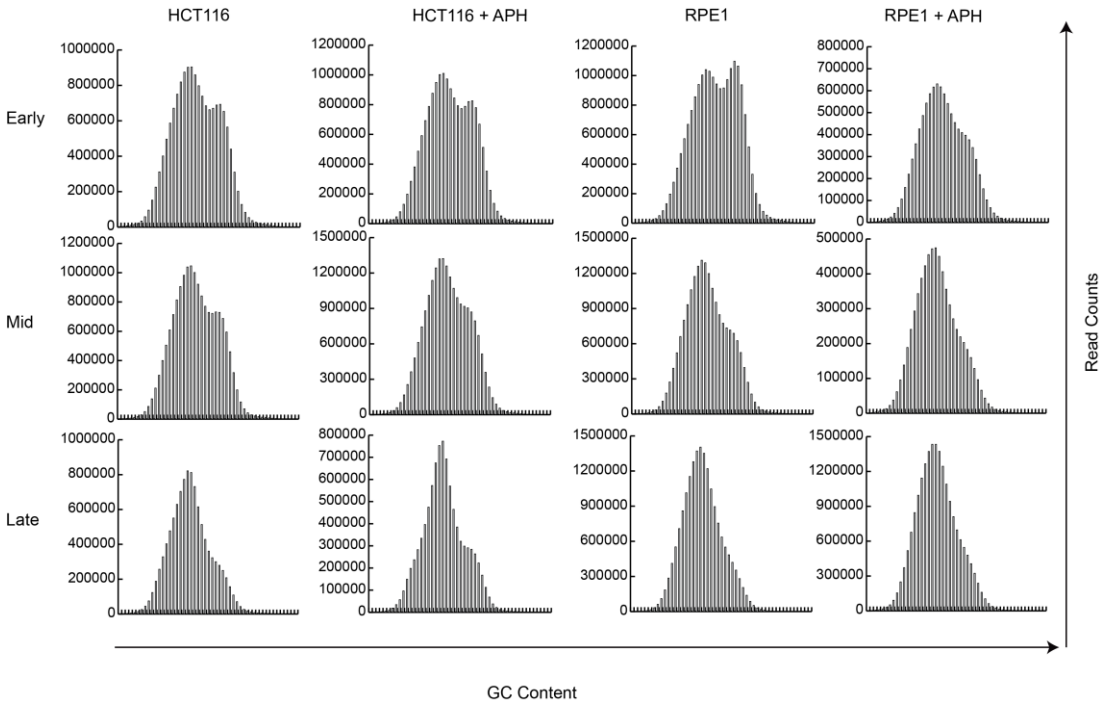


Figure 4-16 GC distribution among Click – seq libraries. Number of reads across 50 GC content bins was calculated using the read_GC function of the RSeQC package. Early samples were found to have a higher proportion of GC-rich reads than late samples.

4.3 Replication timing features in the RPE1 and HCT116 cell lines.

Following the QC analysis of Click-seq data, I wanted to explore the replication timing features of the RPE1 and HCT116 cell lines in unperturbed conditions. I aimed to investigate how the replication timing programmes differed in the two cell types and define the features of the early and late replicating domains within this dataset.

4.3.1 Replication timing profiles in the HCT116 and RPE1 cell line

To explore the replication timing profiles of the two cell types, I calculated the Rvalue in 10,000 bp windows and plotted it across all chromosomes for both HCT116 and RPE1 samples. I found that the Rvalues tended to be similar in neighbouring windows, forming regions of high (tending towards 0.8), low (tending towards 0.2) and medium values, corresponding to putative early, late and transition regions. The Rvalue profiles across chromosomes showed a lot of similarity between the two cell types, but many divergent regions were also observed. Rvalue profiles for the two cell types across chromosomes 18 and 19 are shown in Figure 4-17. Chromosome 18 and 19 are similar in size, but differ substantially in their make up: chromosome 18 is gene-poor and is located predominantly at the nuclear periphery, whilst chromosome 19 is gene-rich and can be found in the interior (Bickmore 2013). Consistently, the replication profile across chromosome 19 included many early replicating regions, in both cell lines, especially in HCT116 (Figure 4-17). The replication profile across chromosome 18, in contrast, was made up of alternating early and late regions. Interestingly, Rvalues in the RPE1 cell line appeared to vary less than in the HCT116 cell line, resulting in a "tighter" replication timing profile. This may be related to differential features of the two cell lines– while HCT116 is of tumour origin, RPE1 is derived from normal tissue through hTERT transformation.

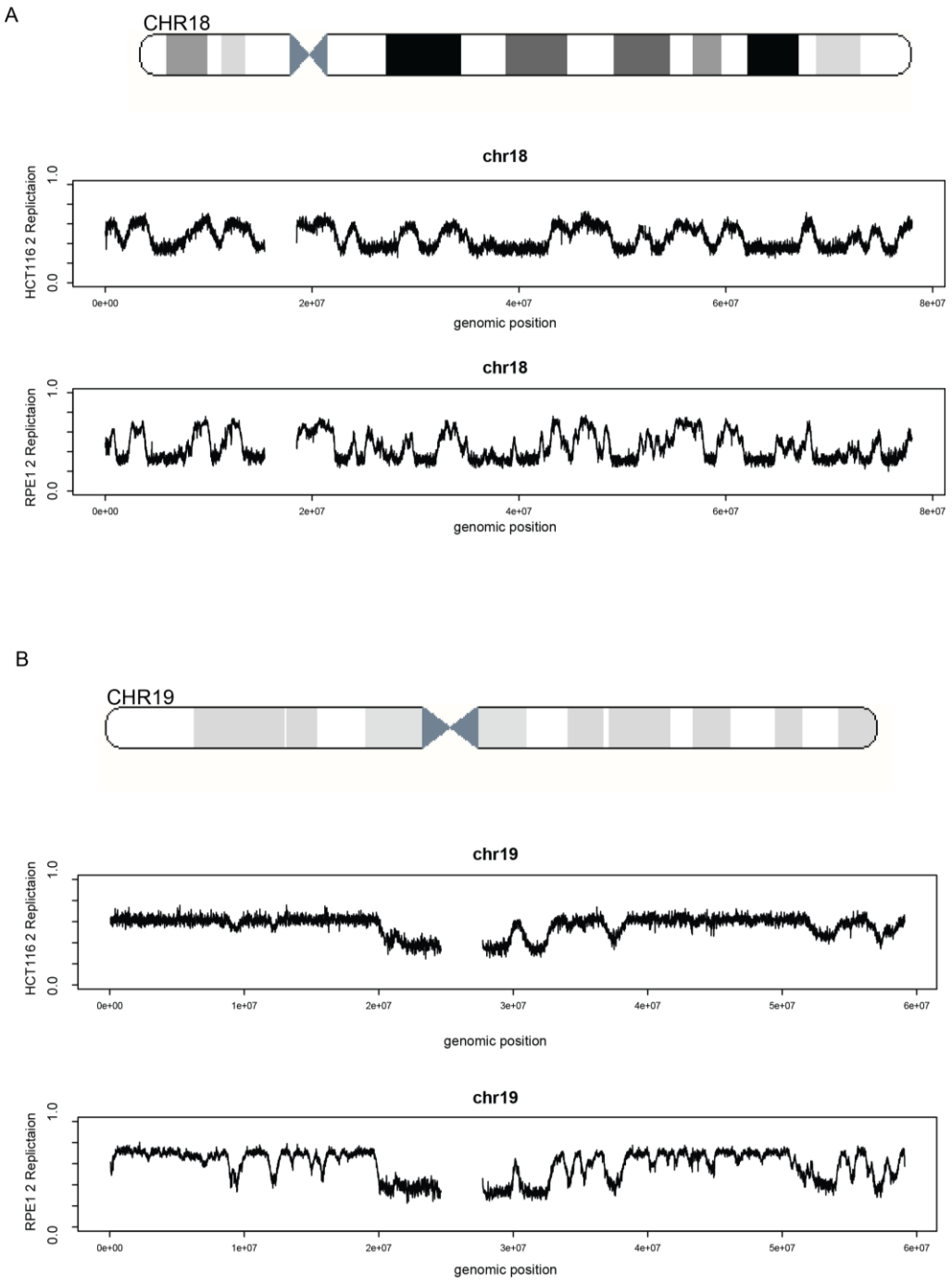


Figure 4-17 Replication timing profiles in the RPE1 and HCT116 cell lines across chromosomes 18 and 19. Rvalues were calculated in 10 kb windows and plotted across chromosomes 18 (A) and 19 (B). The replication timing profile for chromosome 18 showed alternating regions of early and late replication timing that showed a lot of similarity between the two cell types. In contrast, the replication timing profile of the gene-rich chromosome 19 included many early replicating regions, especially in the HCT116 cell line. More differences were observed between the two cell lines for this chromosome.

4.3.2 Partitioning of the genome into replication timing domains

Since plotting replication timing across chromosomes revealed clear delineation of domains with differential replication timing, I set out to partition the genome into defined early, mid and late replicating regions.

Such partitioning has been previously performed for Repli-seq data using different methodologies, including a Hidden Markov Model (HMM) in a 2014 publication (Lubelsky et al. 2014) and the segmentation package DNA Copy, originally developed to define copy number changes in CGH array samples (Ryba et al. 2011; Venkatraman & Olshen 2007). I determined that applying an HMM in this case would be unfeasible, as it would require a proportion of the data with known replication timing to be used as a "training" dataset. In addition, the results would be highly dependent on pre-defined transition probabilities, describing the likelihood of neighbouring windows displaying differential replication timing stages, which are difficult to estimate empirically. I also ruled out using the DNA Copy algorithm, as it is tailored to CGH array derived data in which sharp boundaries corresponding to breakpoints are expected, rather than the smooth transitions observed in replication timing data. Instead, I adapted an "edge filter" approach which has been previously used to define regions of differential supercoiling in the human genome and LADs (Naughton et al. 2013; Guelen et al. 2008). In this approach, an edge filter was applied by calculating the difference in mean Rvalues in 250 1kb windows immediately left and right of a central window, sliding the central window across the chromosome. This resulted in a profile where regions with changes in replication timing corresponded to peaks in the edge filter value. Following a visual comparison of the edge filter and the Rvalue profile across chromosomes, I chose a cut-off value of 0.1 for the edge filter value (shown as a red line in Figure 4-18). Windows with edge filter above 0.3 were considered to span replication timing transition zones, while windows between the edge filter peaks

with values were considered to represent replication domains with stable replication timing.

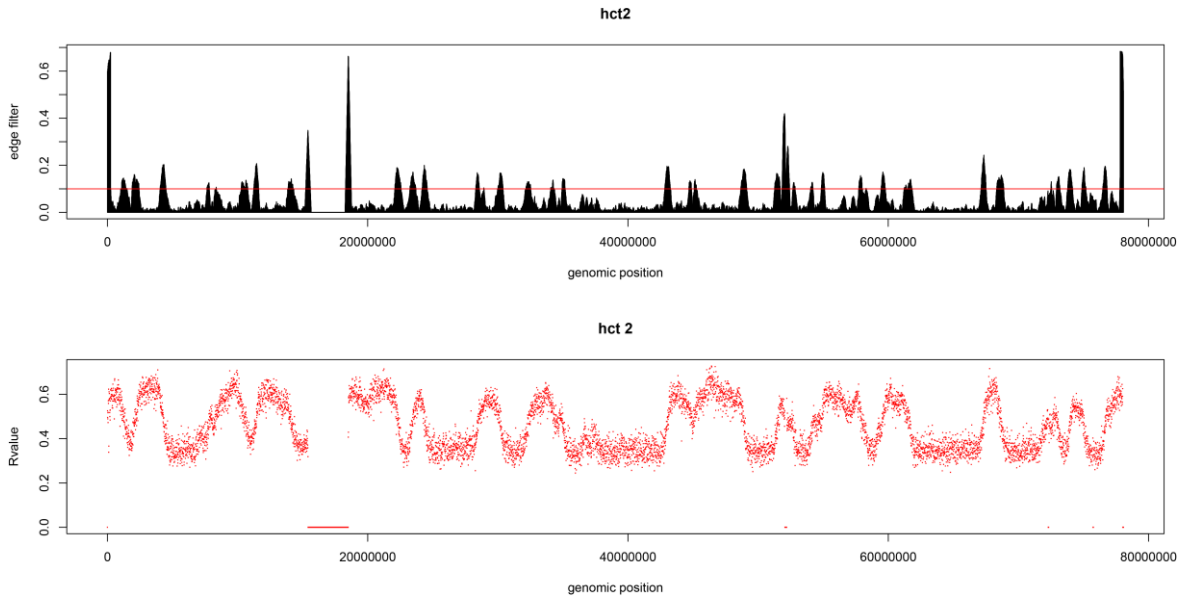


Figure 4-18 Edge filter partitioning of chromosome 3. Edge filter values were calculated across chromosome 3 in sliding windows of 250 x 1 kb. The edge filter values (top) were then compared to the replication timing profile across the chromosome (bottom). Shifts in replication timing corresponded to peaks in the edge filter value. An edge filter values of 0.1 was chosen as a cut off to define replication timing domains (red line across top plot).

Using an edge filter value of 0.1 as a cut off, I selected all the regions contained between two subsequent windows with an edge filter value of 0.1 or more. As replication timing domains are predicted to be at least 30 kb in size, regions smaller than that size were filtered out. Finally, I determined the replication timing of the partitioned domains: first, I calculated the mean Rvalue across all 1000 bp windows encompassed within a domain; mean Rvalues across all domains and all samples were then clustered using kmeans clustering and assigned either to an early, mid or late cluster. A boxplot of the distributions of mean domain Rvalues within each cluster is shown in Figure 4-19. The early cluster contained 40.4 % of all identified domains, the late cluster- 30.6 % and the mean cluster-29 %.

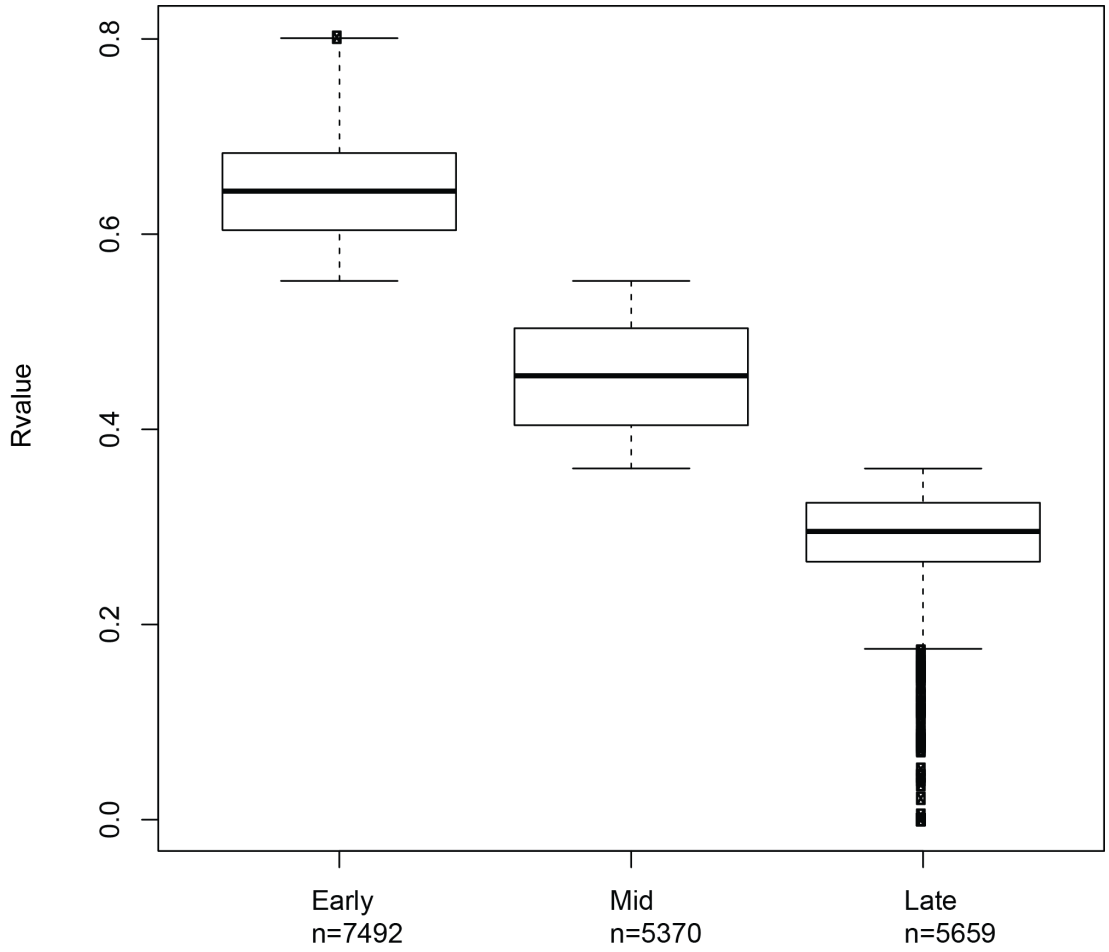


Figure 4-19 Distribution of mean domain Rvalues across the early, mid and late clusters defined by k-means clustering. Numbers of domains within each cluster is shown along the x-axis labels.

A visual comparison of the partitioned domains to the raw data in the HCT116 sample is represented in Figure 4-20. It showed that the edge filter correctly assigned regions with high Rvalues and high density of early reads as early domains and regions with low R values and predominantly late reads as late domains.

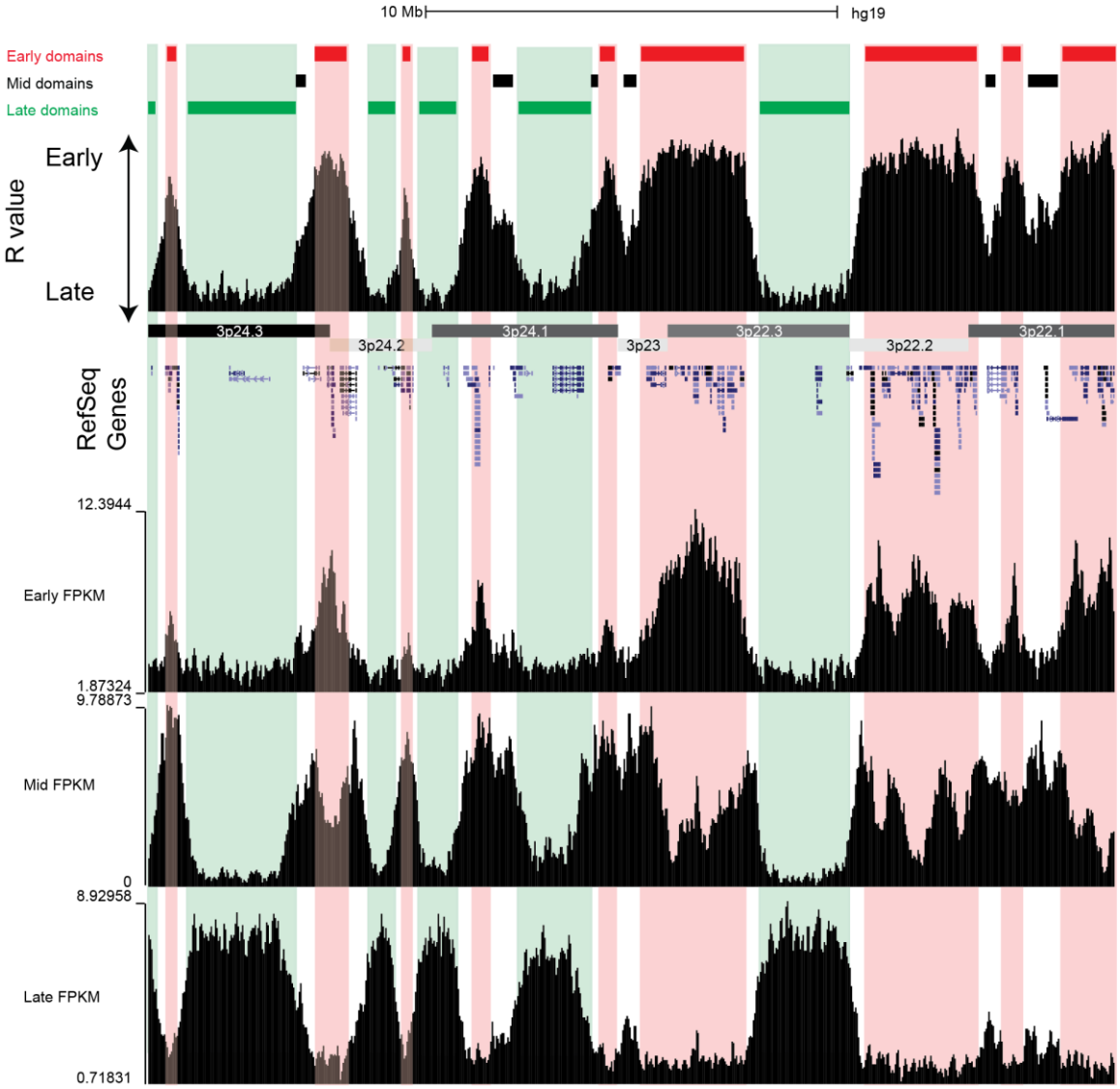
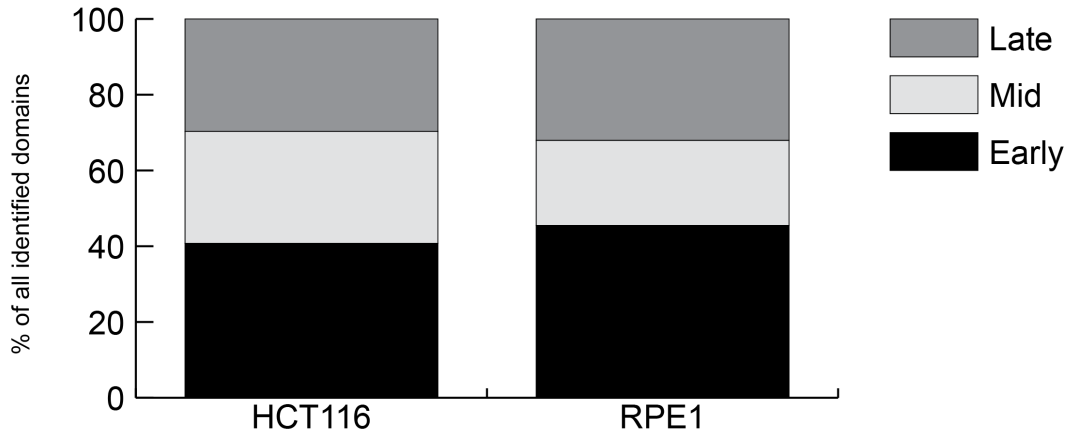


Figure 4-20 Comparison of partitioned domains to raw data. A 20 Mb region on the p arm of chromosome 3 is presented along with data for the HCT116 cell line. The defined domains are shown as blocks at the top of the figure: early domains are shown as red blocks, mid domains are in black and late domains are shown in green. The regions corresponding to the early and late domains are shaded in red and green respectively. R values in 1000 bp windows across the genome are shown below the domains, while raw FPKM values from the early, mid and late samples are shown in the lower part of the figure.

4.3.3 Replication domain features in the RPE1 and HCT116 cell line

Following the edge filter analysis, a total of 1769 domains were identified in the HCT116 cell line and 2810 in the RPE1 cell line. The two cell types showed different numbers of domains designated as early replicating (40% in the HCT116 cells and 45% in RPE1), and showed different distributions of mid and late domains (Figure 4-21): HCT116 cells appeared to have similar proportions of mid and late domains, while regions designated as late were more prevalent in the RPE1 cell type. Taking into account the size of each identified domain, I also assessed the proportion of the genome classified as early, mid or late in each cell line. I found that a smaller proportion of the genome was covered by the defined domains in the RPE1 cell type, which may suggest that this cell type has more transition zones between domains. Within the HCT116 cell line, around 45% of the genome was identified as early replicating and 31% as late replicating, while in RPE1 the corresponding numbers were 29% and 38%. The share of mid-replicating sequences was much lower: only 11% in HCT116 cell lines and 6% in RPE1, indicating that mid-replicating domains are smaller in size. To verify this, I plotted the size distribution of early, mid and late domains within each cell line (Figure 4-22). I found that, as indicated by the previous analysis, mid domains were smaller than early and late domains in both cell lines. Surprisingly, the domain sizes in the RPE1 cell line appeared smaller than in the HCT116 cell line. A visual inspection of the partitioning against the R-values across the chromosomes indicates that large early and late replication timing zones in the RPE1 cell type are more punctuated and less contiguous than in HCT116 (can be observed in Figure 4-17). While differences in replication timing between different lineages are well described, variations in domain structure and size have not been characterised in detail. One possible explanation is that the tumour origin of HCT116 affects the temporal control of origin firing, resulting in a more variable domain size.

A



B

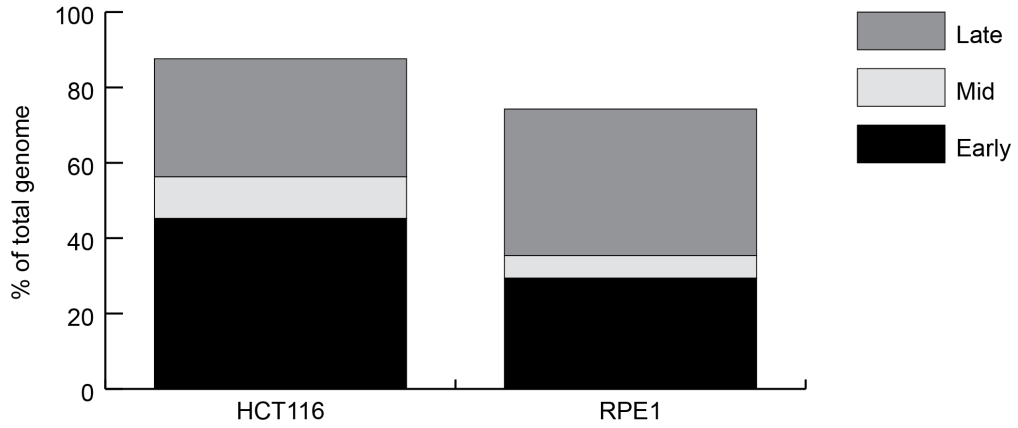


Figure 4-21 Domain distribution in the RPE1 and HCT116 cell lines. A. Percentage of domains in the early, mid and late category within the two cell types. B. Proportion of the total genomic sequence covered by early, mid and late domains in the two cell types. Remainder of the genome not covered by the domains is made up of transition zones and regions that cannot be mapped and sequenced, such as centromeres and other repetitive sequences.

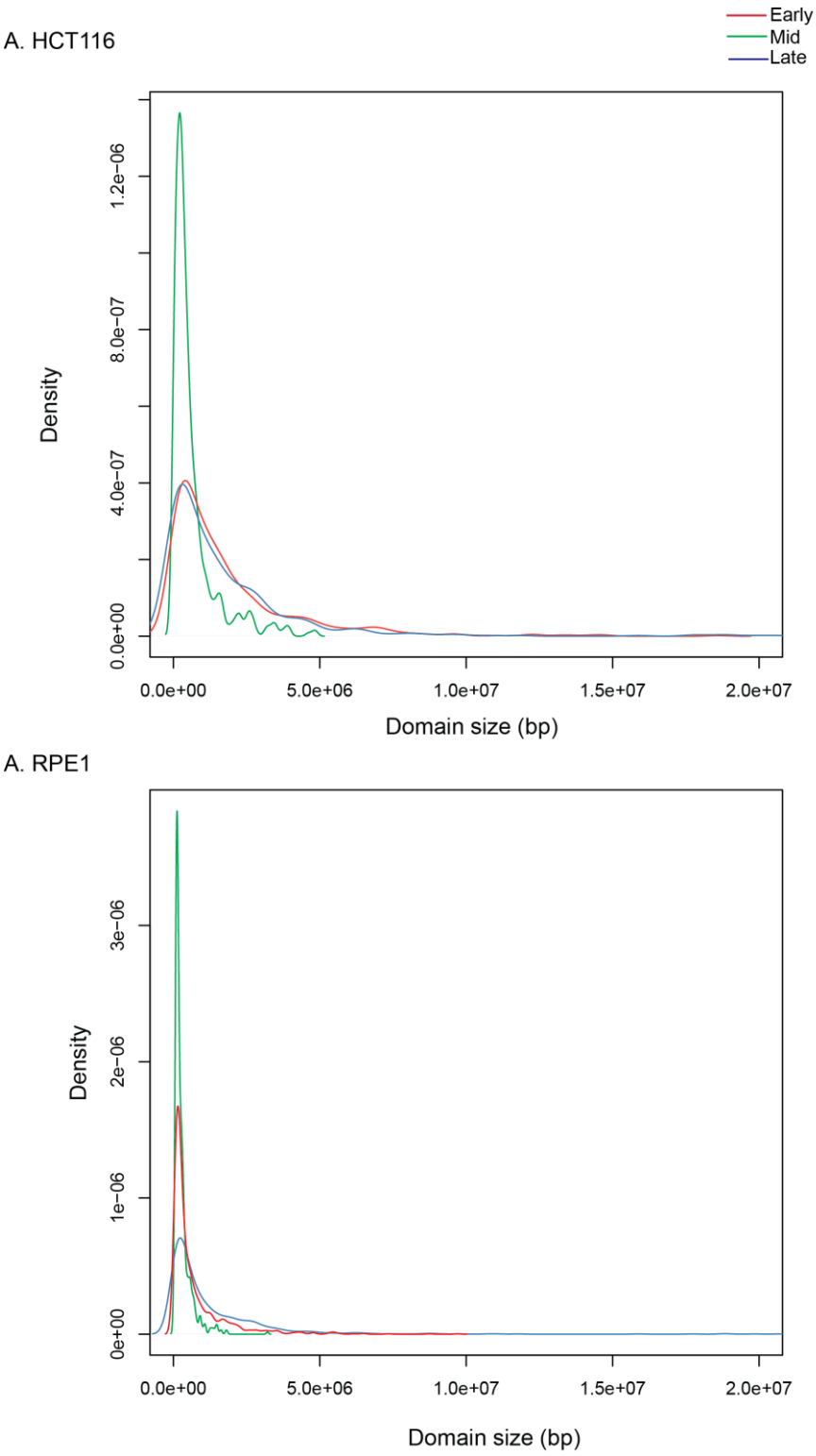


Figure 4-22 Distribution of domain sizes in the RPE1 and the HCT116 cell lines. Frequencies of domain sizes were plotted for the HCT116 cell line (A) and the RPE1 cell line (B). Frequency of early domains is shown in red, mid in green and late in blue.

I next examined GC content across the different domain categories. Previous replication timing studies have shown that GC-rich regions of the genome tend to replicate early, while GC-poor regions replicate later (Hansen et al. 2010). As expected, I found a similar trend in both cell lines: on average, early replicating domains were made up of more GC-rich sequences than mid and late domains (Figure 4-23).

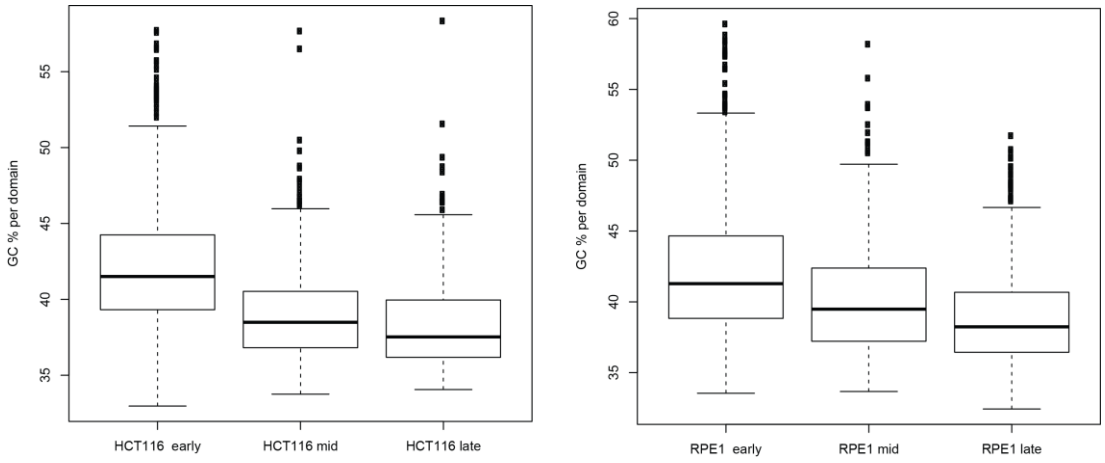


Figure 4-23 GC content across different domains in the HCT116 and RPE1 cell lines. GC% was plotted for each domain within the early, mid and late category in the HCT116 (A) and the RPE1 (B) cell lines.

I also investigated the gene overlap end expression rates across the different domains in the two cell types (Figure 4-24). First, I assessed the numbers of RefSeq genes contained within each domain type in each cell type. As expected, the early domains contained the majority of genes in both cell types. In both cell types, the late domains contained more genes than the mid domains, likely due to the fact that mid domains comprised a much smaller proportion of the genome than late ones. When only genes expressed in each cell line were considered, a similar relationship was observed: the overwhelming majority of expressed genes were contained within early domains, while mid and late domains contained a significantly smaller number of active genes. Using the RNA-seq data described in

Chapter 3.4, I compared the FPKM distributions for genes located within early, mid and late domains within the two cell types. Unsurprisingly, I found that genes contained within early domains showed the highest FPKM values, while genes contained in late domains showed the lowest.

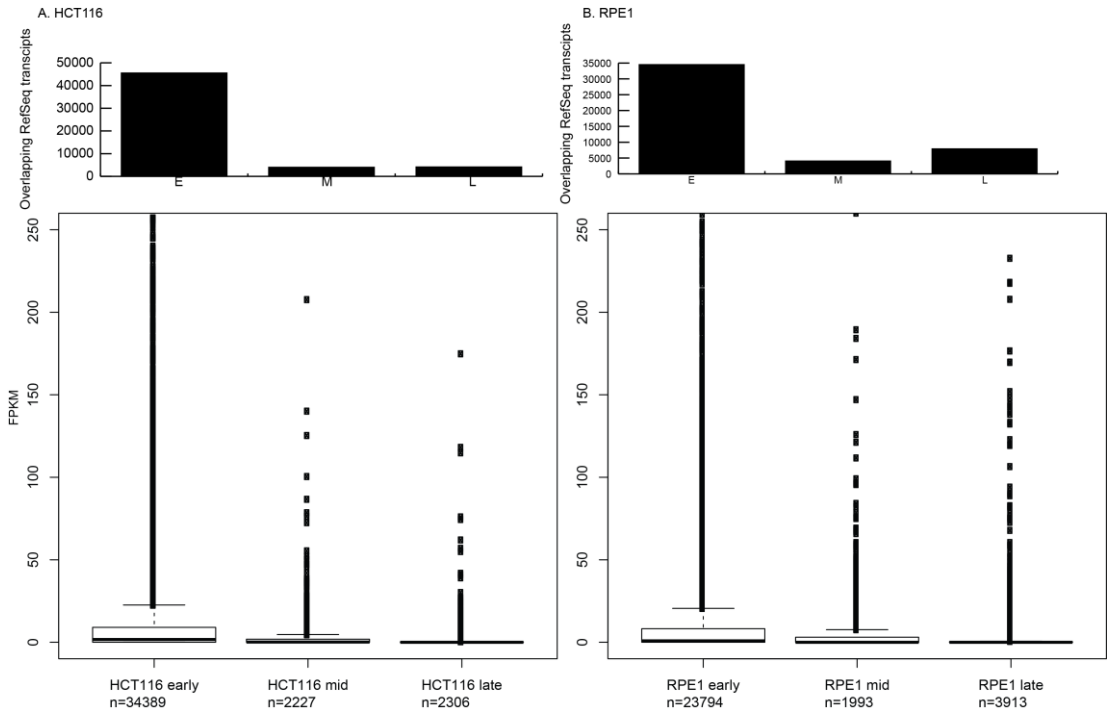


Figure 4-24 Gene density and expression level across different domains in HCT116 and RPE1 cell lines. Top histograms show the number of RefSeq genes contained within early (E), mid (M) and late (L) domains in each cell type. Bottom boxplots show the distribution of FPKM values from the RNA-seq data discussed in Section 3.4 for transcripts located within the early, mid and late domains in each cell type. Labels below boxplots denote the number of transcripts for each domain type.

Finally, I compared the domain overlap between the two cell types. Previous comparisons between distinct cell types showed that around 50% of the genome has a constant replication timing (Ryba et al. 2010). An intersection of the early domains in the two cell types resulted in 881 overlapping early domains, corresponding to 22% of the genome. For the mid domains, 151, corresponding to just 1.65% of the genome were shared between the two cell lines and when late domains were considered, I found that 205 overlapped, containing 13% of the

genome. Overall, 36% of the whole genome appeared to have identical replication timing in the two cell types

4.4 Effect of replication stress on replication timing in the RPE1 and HCT116 cell lines

"Replication stress" is a broad term which covers a range of conditions that interfere with the normal progression of DNA replication. Outcomes of replication stress include a slowdown in fork speed, fork stalling, accumulation of DNA damage and even structural genomic instability (Burrell et al. 2013; Chan et al. 2009; Gaillard et al. 2015). Induction of replication stress has been described under many conditions: physiological replication stress has been observed upon oncogene activation and entry into S-phase with an insufficient nucleoside pool (Minocherhomji et al. 2015; Bester et al. 2011); endogenous replication stress has been described in CIN^+ colorectal cancer lines (Burrell et al. 2013). Pharmacologically, it can be triggered by drugs such as aphidicolin and hydroxyurea. Aphidicolin, used to induce CFS formation, is an inhibitor of DNA polymerases, predominantly DNA polymerase α, which works by blocking dCTP incorporation by polymerase (Krokan et al. 1981; Baranovskiy et al. 2014). Hydroxyurea functions through a different mechanism, by depleting the dNTP pool inside cells, which results in fork stalling and ultimately DSBs (Petermann et al. 2010). It is not known how aphidicolin leads to fork stalling or inactivation, however, it is well known that high concentrations of the drug can completely inhibit replication.

Surprisingly, genome-wide changes in replication timing have never been studied under conditions of replication stress. DNA combing experiments have shown that over-expression of Myc causes premature origin firing and an increase in origin density (Srinivasan et al. 2013) and it is well known that aphidicolin in particular and replication stress in general can cause recruitment of additional origins to allow replication to proceed on time (Letessier et al. 2011; Blow et al. 2011). However, it is not known how these cellular responses to replication stress affect the genome wide replication timing programme. To investigate this, I analysed Click-seq data

generated from RPE1 and HCT116 cells exposed to replication stress caused by low-dose aphidicolin.

4.4.1 Replication timing profiles in the HCT116 and RPE1 cell lines under conditions of replication stress.

Low-dose aphidicolin treatment of asynchronously dividing HCT116 and RPE1 cells resulted in a pronounced accumulation of cells in S-phase, illustrating the inhibitory effect of the drug on replication dynamics (Figure 4-25). The increase in S-phase cells suggests that replication is slowed down upon aphidicolin treatment. However, this may indicate either a delay affecting the whole genome, or a more-locus specific effect, with certain locations showing increased susceptibility to changes in replication timing in the presence of aphidicolin. A genome-wide delay in replication timing, affecting all loci in a similar manner, would result in a replication timing pattern which would be undistinguishable from the control population. On the other hand, if aphidicolin induces locus –specific effects, they could be observed as regions of differential replication timing between the control and the drug-treated population. The initial assessment of read density profiles indicated that changes can be observed between the control and drug treated samples (Figures 4-10 and 4-11).

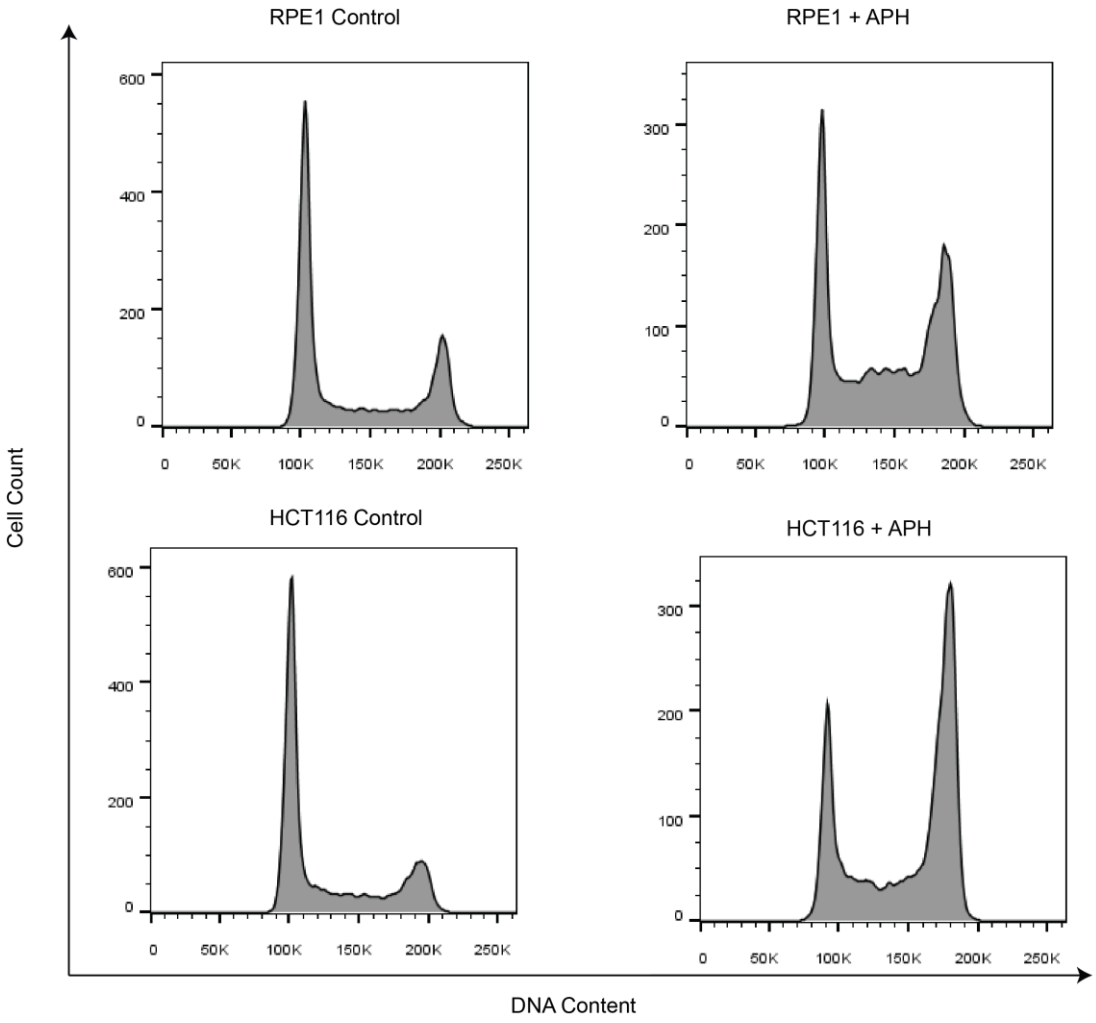


Figure 4-25 Effects of aphidicolin on the cell cycle profile on RPE1 and HCT116 cells. Treatment with low dose of aphidicolin (0.4μM) for 24 hours was found to cause an accumulation of S-phase cells in both cell types, as indicated by a change in the FACS profile of PI-stained cell populations.

To explore the extent of these changes, I plotted Rvalues for control and aphidicolin treated samples in 10,000 kb windows across different chromosomes and assessed the changes induced by aphidicolin. In the HCT116 cell line I found that aphidicolin induced subtle changes in the replication timing profile (Figure 4-26). Surprisingly, the induced changes were not uni-directional towards later replication; instead, both changes from an early to later and from a late to earlier replication timing were observed across the chromosome.

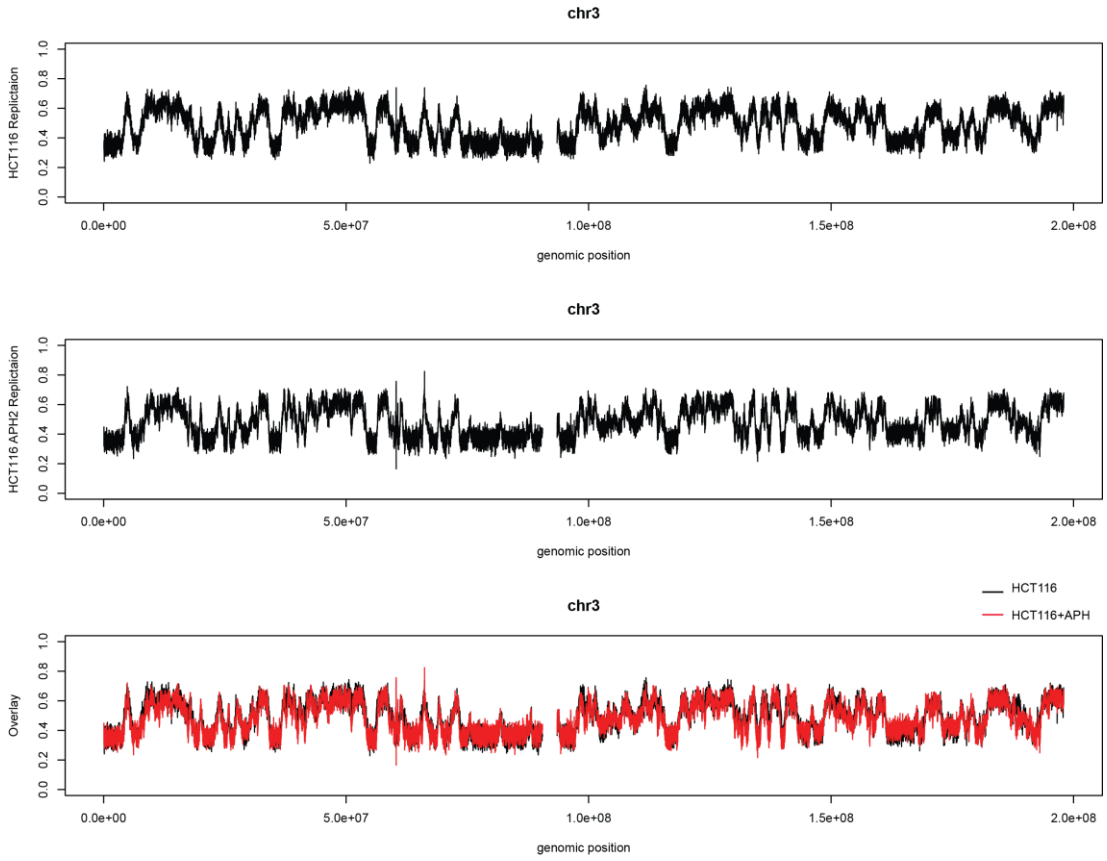


Figure 4-26 Effect of aphidicolin treatment on replication timing across chromosome 3 in HCT116 cells. Rvalues in 10 kb windows were plotted across chromosome 3 for the HCT116 control sample (top plot) and the HCT116 aphidicolin-treated sample (middle plot). An overlay of the two values is shown in the bottom plot, with control Rvalues shown in black and aphidicolin-treated Rvalues shown in red. Changes from both higher to lower R value (indicating an early to later shift) and lower-to-higher R value (indicating a late to earlier shift) can be observed.

Next, I wanted to assess if similar changes could be observed across chromosome 3 in the RPE1 cell line. Unlike the HCT116 cell line, where one of the biological

replicates was of insufficient quality, I could investigate if any observed changes were consistent across the biological replicates. An immediate observation for this cell line was that aphidicolin treatment resulted in a much “noiser” replication timing profile, indicating that replication stress caused some mis-regulation of replication progression (Figure 4-27). In addition, I observed similar effects upon aphidicolin treatment as in the HCT116 cell line: again, subtle bi-directional changes could be observed, with both regions transitioning from early to later replication timing and the opposite. Strikingly, a comparison between the two biological replicates revealed that these changes occurred at similar genomic locations across the replicates, suggesting that aphidicolin treatment may result in recurrent changes in replication timing at particular genomic locations.

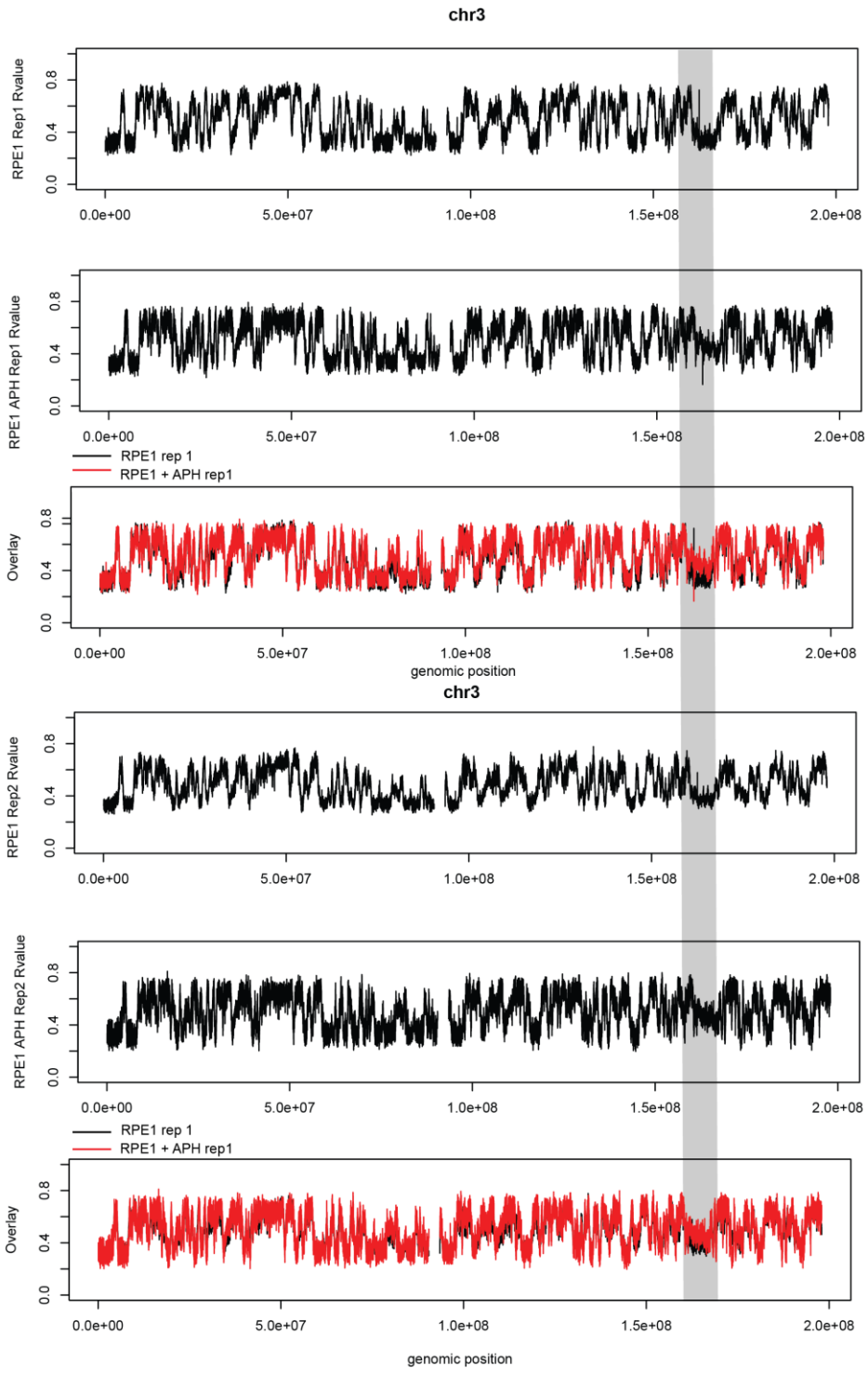


Figure 4-27 Effect of aphidicolin treatment on replication timing across chromosome 3 in the RPE1 cell line. Rvalues in 10 kb windows plotted across chromosome 3 for the two control sample replicates and the two HCT116 aphidicolin-treated replicates. An overlay of the two datasets is shown for each replicate, with control Rvalues in black, and aphidicolin-treated Rvalues in red. Similar changes across the biological replicates can be observed, such as in the region highlighted in grey.

In addition to chromosome 3, I also concentrated on chromosome 18 and 19, which were observed to show very different replication timing profiles (Chapter 4.4.1, Figure 4-17), consistent with the differences in GC composition, gene density and nuclear positioning between these two chromosomes. In both cell types, the early-replicating chromosome 19 showed an overall change towards later replication timing. In contrast, across chromosome 18, early replicating regions appeared to shift to a later timing while late regions changed towards earlier values – a tendency that was especially clear in the HCT116 cell line (Figure 4-28). In conclusion, my observations across the different cell types and chromosomes suggested that rather than a universal delay, replication stress induces bi-directional shifts in replication timing.

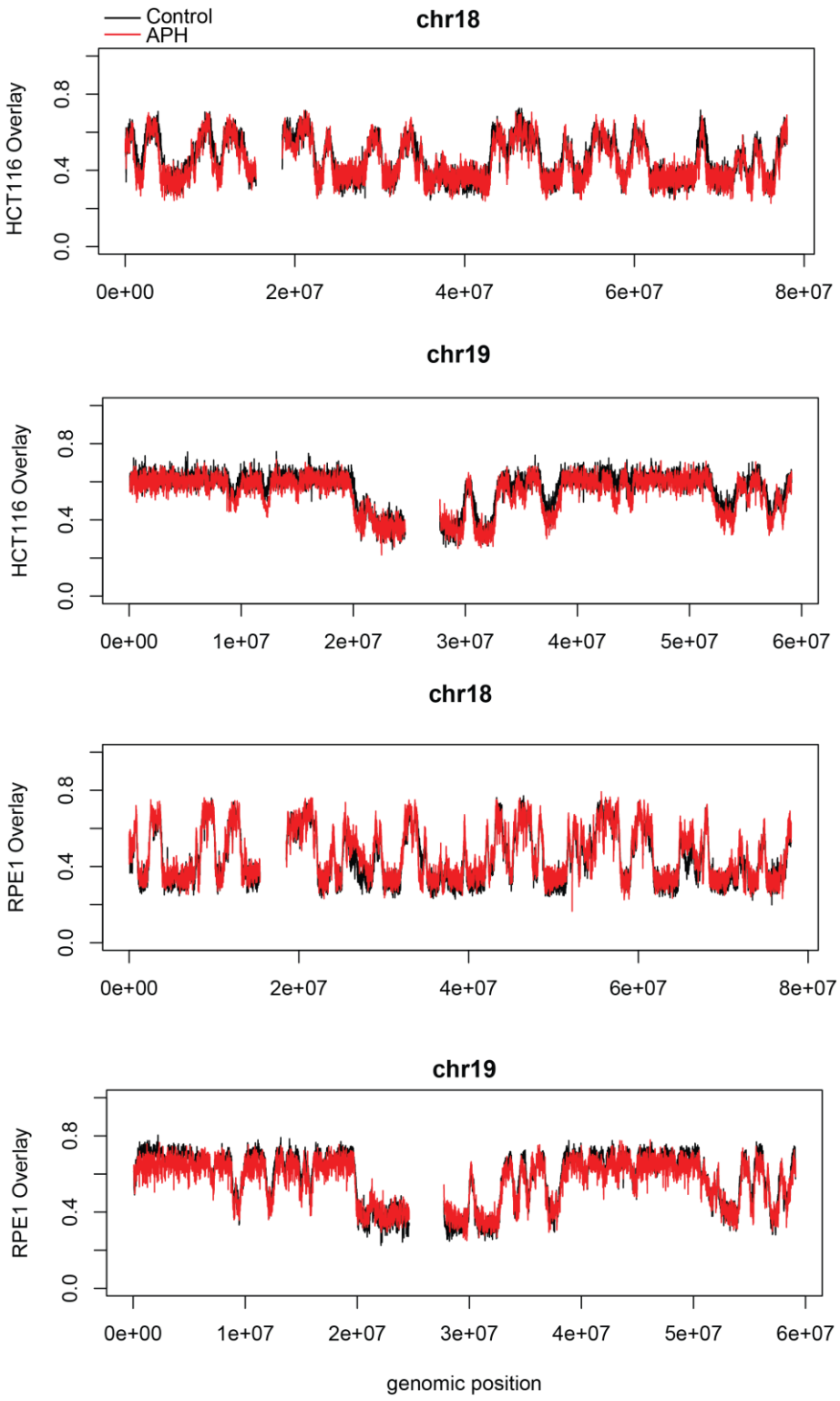


Figure 4-28 Replication stress induced replication timing changes across chromosomes 18 and 19 in RPE1 and the HCT116 cell lines. Overlay of the Rvalues in 10,000 bp windows is shown for each chromosome and each cell type, with control values shown in black and values for the aphidicolin treated sample shown in red.

To further examine the nature of the changes, I constructed heatmaps of Rvalues calculated in 10 kb windows for the two cell lines, comparing controls and aphidicolin-treated samples (Figure 4-29). Confirming the data shown in Figures 4-26, 4-27 and 4-28, the heatmaps indicated bi-directional changes in the aphidicolin-treated samples, with some windows changing to an earlier replication and some windows-to later timing. Very few sharp changes could be seen, with most windows changing to a slightly earlier or slightly later timing in the presence of replication stress.

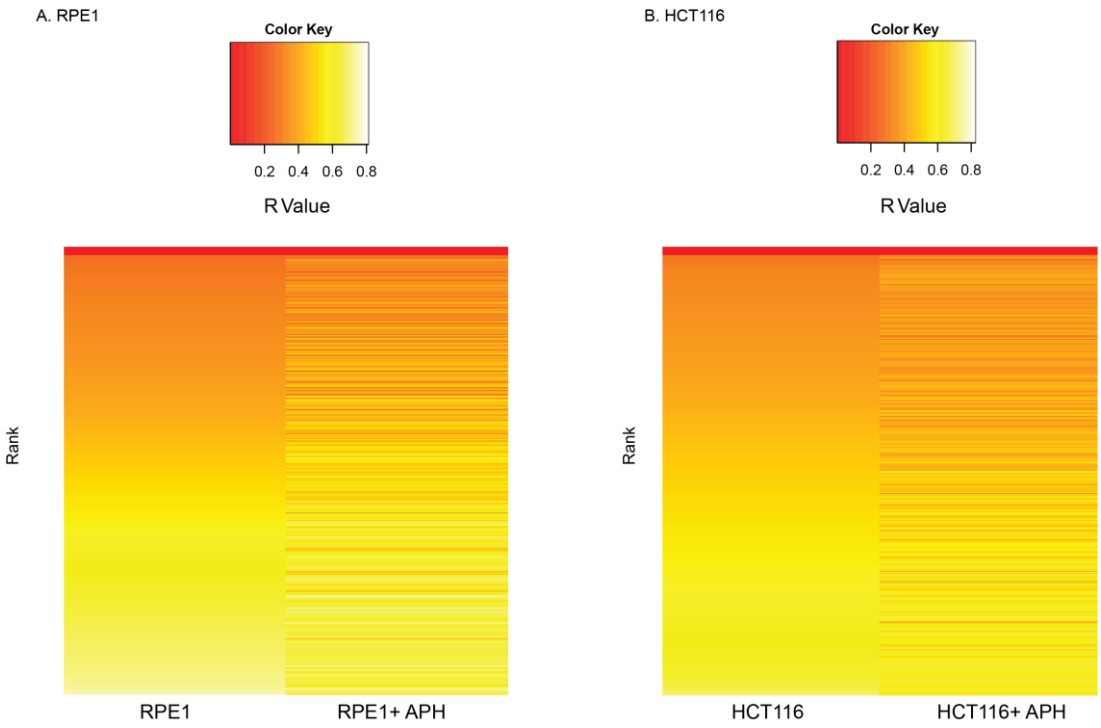


Figure 4-29 Rvalue changes in aphidicolin treated samples across chromosome 3. Rvalues in control RPE1 (A) and HCT116 (B) cell lines were calculated in 10,000 kb windows and ranked from lowest to highest. The corresponding Rvalue for each window in the aphidicolin treated sample is shown next to the control.

4.4.2 Replication timing domain changes upon replication stress.

As Rvalues indicated subtle changes in replication timing in the presence of replication stress, I next explored how the aphidicolin treatment affected the domain structure within the two cell types. I segmented the genome in the aphidicolin treated samples following the method described in Section 4.4.2 and examined the differences in domain characteristics in the control and the aphidicolin-treated samples. Within the HCT116 cell line, I found that the number of identified domains had increased from 1769 to 2462; an increase was also observed in the RPE1 cells, from 2810 to 3347. In both cell types, while the total number of identified domains was increased, the proportions of early, mid and late domains stayed similar in the aphidicolin-treated and the control sample (Figure 4-30). In HCT116 cells, the proportion of the genome contained within early domains was decreased in the aphidicolin treated sample (from 45% to 29%), while the proportion of the genome contained within late domains was increased (from 31% in controls to 39% in the drug-treated sample). In RPE1 cells, no big changes were seen in the proportion of genomic sequence designated for each of the domain categories.

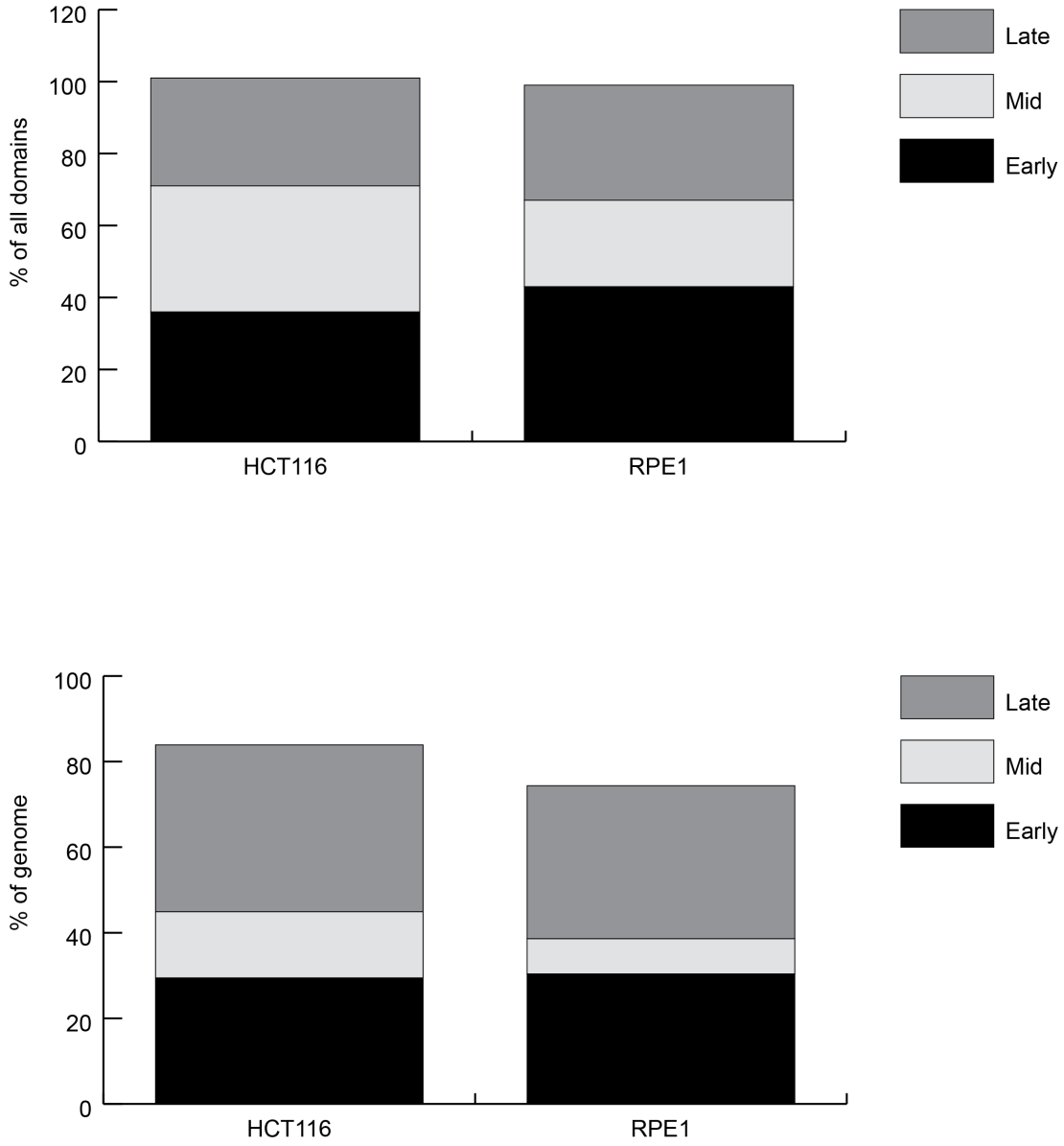


Figure 4-30 Domain distribution in aphidicolin treated RPE1 and HCT116 cells. Top graph shows the percentage of domains in the early, mid and late category within the two cell types. Bottom graph shows the proportion of the total genomic sequence covered by early, mid and late domains in the two cell types.

I next investigated whether the size of domains changed upon replication stress induction (Figure 4-31). In both cell types, the domain size was decreased in the presence of aphidicolin; this reduction in size was particularly pronounced for the early and mid domains. The decrease could be due to reduced fork speed in the aphidicolin samples, causing forks to travel smaller distances away from initiation zones.

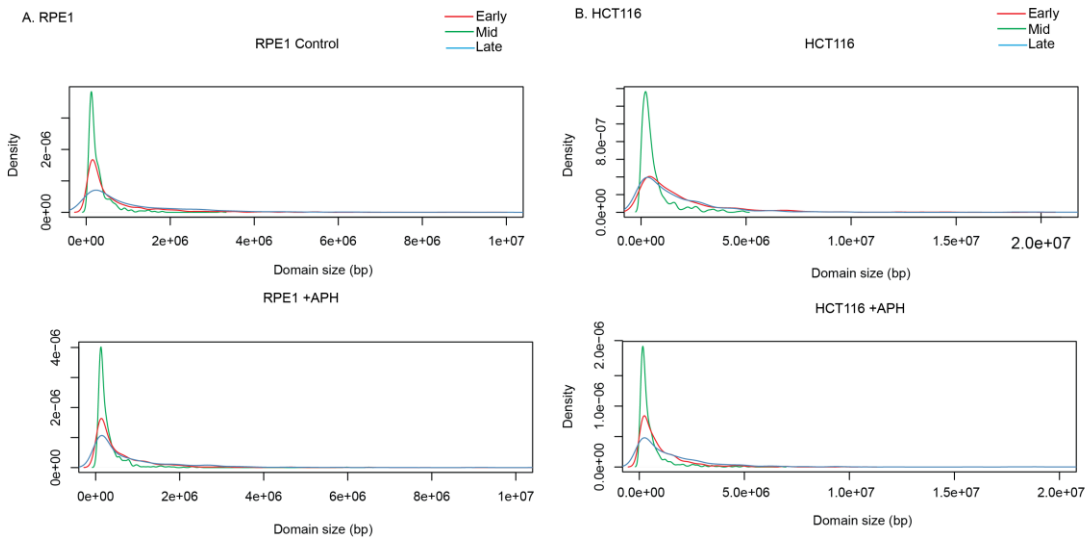


Figure 4-31 Changes in the distribution of domain sizes in the RPE1 (A) and the HCT116 (B) cell lines in response to replication stress. Frequency distributions of domain sizes are shown, with early shown in pink, mid in turquoise and late in blue.

I also investigated how GC composition of domains changed upon replication stress induction (Figure 4-32). As described in section 4.4.2, under control conditions, I found that early domains had higher average GC content than mid and late domains in both cell types. This trend could also be observed in both cell lines under conditions of replication stress, however an increase in the GC content of early, mid and late domains could be clearly observed in the HCT116 cell line. Late domains in the RPE1 cell line also showed a small increase in GC content in the presence of aphidicolin compared to the control sample.

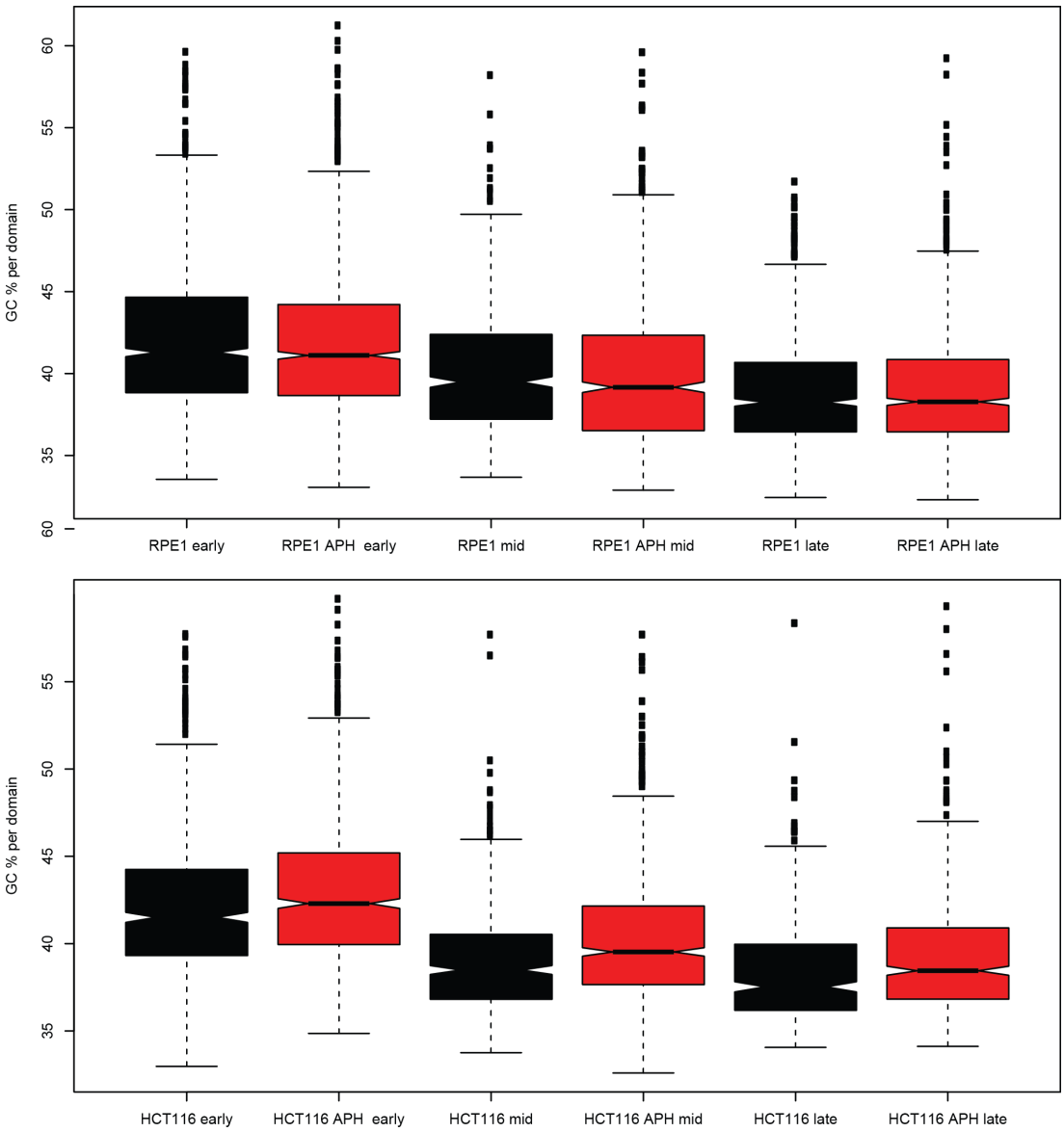


Figure 4-32 GC content across different domains in the RPE1 (top) and HCT116 (bottom) cell lines under control conditions and in the presence of aphidicolin. GC% is plotted for each domain within the early, mid and late category.

There were also differences in the gene density between domains following induction in replications tress. In HCT116 cells, early domains overlapped with 20% fewer RefSeq genes in the aphidicolin sample compared to the control. In both cell types, late domains overlapped with an increased number of genes in the samples treated with aphidicolin: around 60% more RefSeq genes were contained within the late domains in the aphidicolin samples compared to the controls (numbers

presented in Figure 4-33). When only genes expressed in each cell lines were considered, this trend was preserved: early domains in the HCT116 cell line overlapped with 34,389 expressed genes in control cells and just 23,053 expressed genes in the aphidicolin-treated sample. Late domains in HCT116 cells overlapped with 3657 expressed genes in the aphidicolin treated sample and 2305 in the control cells. In conclusion, replication stress causes a disruption in the replication timing program which causes a subtle loss of features normally associated with early and late replication timing.

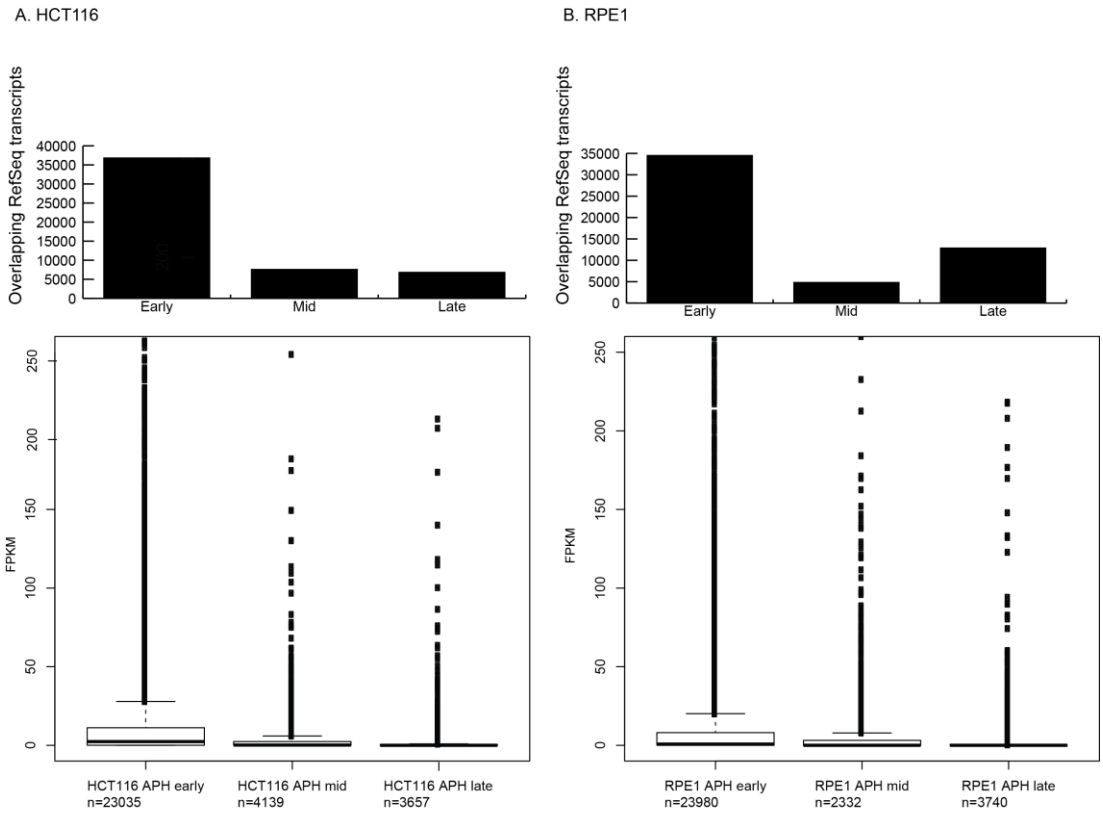


Figure 4-33 Gene density and expression rates across different domains in HCT116 and RPE1 cell lines following aphidicolin treatment. Top histograms show the number of RefSeq Genes contained within early (E), mid (M) and late (L) domains in each cell type following aphidicolin treatment. Bottom boxplots show the distribution of FPKM values for genes located within the early, mid and late domains for each cell type following aphidicolin treatment. Labels below boxplots denote the number of transcripts for each domain class.

Finally, I assessed the extent of domain overlap between the control and aphidicolin treated sample in each cell line. Within the HCT116 cell line, I found that 742 early domains retained their early replication timing following aphidicolin treatment, comprising 24% of the genomic sequence. In contrast, only 139 late domains, comprising just 2.17% of the genomic sequence retained late replication timing following aphidicolin treatment. . In the RPE1 cell line, 386 early domains retained their replication timing under replication stress, covering 7.87% of the genome. Only a small fraction of the genome, less than 0.3%, retained late replication timing in this cell type. In both cell types, very small proportions of the genome changed from early to mid and late to mid, or showed more extreme changes such as early to late and the opposite. Curiously, none of the small number of domains showing extreme changes in replication timing mapped to active CFS regions.

4.5 Replication timing and CFS instability

Late replication timing is considered to be one of the defining characteristics of CFS regions. Initial FISH-based experiments were performed across FRA3B, FRA7H and FRA6E and indicated that they span regions which replicated late and were delayed in the presence of aphidicolin (Wang et al. 1999; Palumbo et al. 2010; Hellman et al. 2000). The late replication timing of FRA3B was later confirmed with results by SNS fragment isolation (Palakodeti et al. 2009) and by Repli-seq (Letessier et al. 2011). Due to the cell type specificity of CFS regions, only two studies to date have tried to correlate CFS expression to Repli-seq generated tissue-specific replication timing profiles, with results suggesting that CFS regions replicate later and span fewer origins in the cell types in which they are fragile (Le Tallec et al. 2011; Letessier et al. 2011). The effect of aphidicolin on replication speed across the genome and at the FRA3B locus has been studied through the DNA fibre FISH technique, which revealed that aphidicolin caused a slow-down in fork speed across the genome, which was not more severe at the FRA3B site. However, in-depth, cell type matched studies of how aphidicolin affects the replication timing programme across CFS regions have never been performed. With my Click-seq dataset, I was able to define

for the first time the replication timing features across active and inactive CFSs in unperturbed cells and in the presence of aphidicolin.

4.5.1 Replication timing across CFS regions

To investigate the replication timing features of CFSs, I plotted the distribution of FPKM values from the early, mid and late fractions in 10,000 kb windows across the sites I identified in Chapter 3.1.

4.5.1.1 Replication timing across fragile locations in the RPE1 cell line

I first examined the replication landscape across FRA1C at chromosome 1p31.2, the most fragile location in the RPE1 cell line, which harboured 18.6 % of all CFS breaks (Figure 4-34). In both the RPE1 and HCT116 cell line, this location appeared to be replicated from long travelling forks originating from early initiation zones located from 0.5 Mb to 1 Mb away. The fragile location was replicated late in both cell types in control cells and also in the presence of aphidicolin. Unfortunately, it was not possible to determine if the site was passively replicated by forks converging from long distance or if late origins were activated instead. Surprisingly, a delay in replication across FRA1C was not observed upon aphidicolin treatment. On the contrary, the origins surrounding the site, where forks originated, appeared to shift from a mid to an early timing in conditions of replication stress. The replication landscape at this region did not differ substantially between RPE1 and the HCT116 cell line, where breaks at this CFS accounted for only 5.8 % of all breaks. In the unperturbed HCT116 population, a smaller proportion of the region appeared to be replicated late, suggesting that forks may be able to move faster across FRA1C in this cell line. Like in the RPE1 cell line, the origins surrounding FRA1C appeared to shift from a mid to an early replication timing upon aphidicolin treatment.

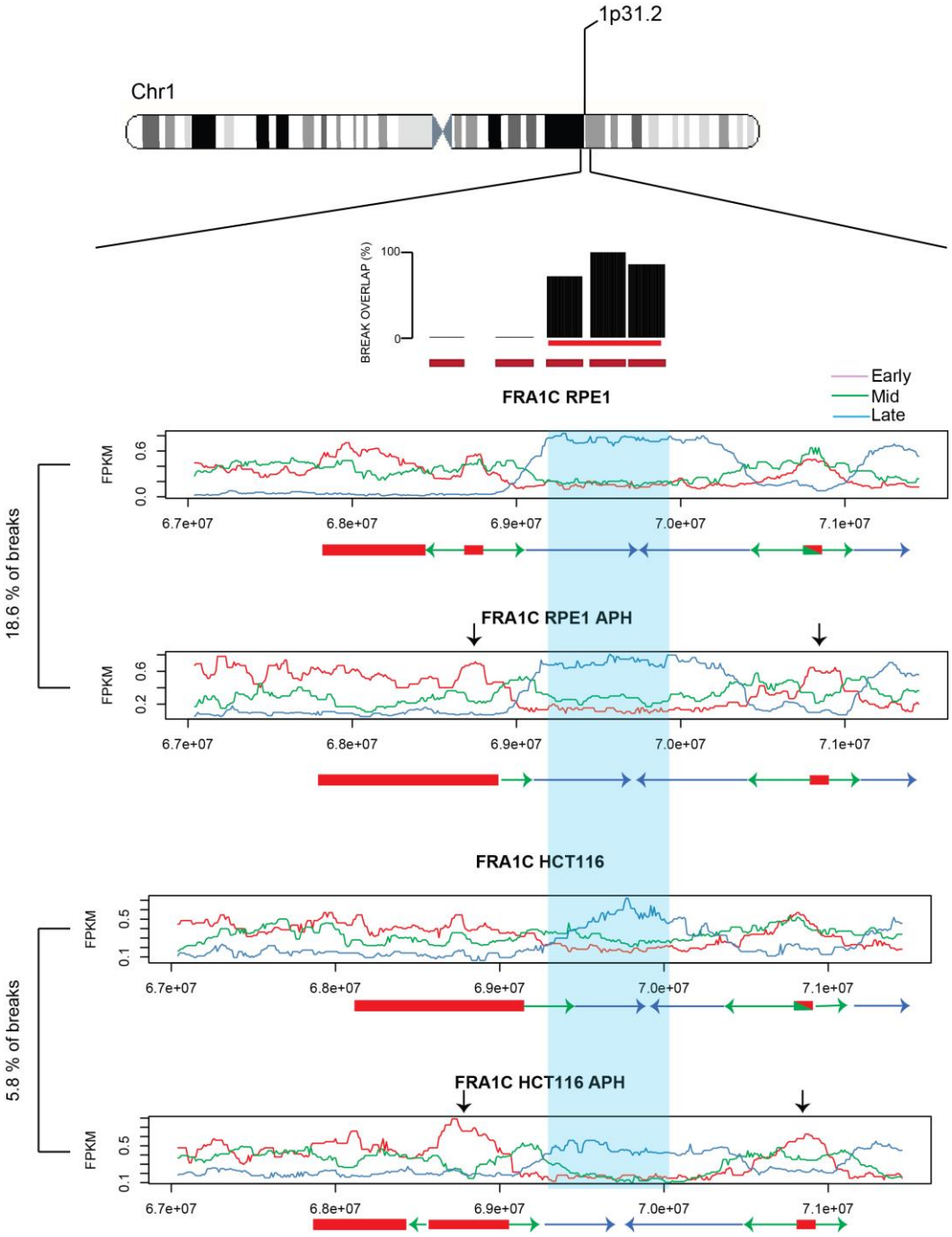


Figure 4-34 Replication landscape at the FRA1C site. FPKM read density in 1000 bp windows are presented for each early (red), mid (green) and late (blue) fractions across FRA1C for the two cell lines in unperturbed conditions and upon aphidicolin treatment. The fragile region is shaded in blue and the locations of the BAC probes used for fine-mapping FRA1C (discussed in Section 3.2.1.1) are shown at the top, with the corresponding break overlap for each probe. Underneath the tracks, a schematic diagram of replication dynamics is drawn with suspected initiation zones represented as rectangles and travelling forks represented as arrows. Black vertical arrows denote initiation zones which fire earlier in the presence of aphidicolin.

The other fine-mapped location in RPE1 cells, the novel 4q32.2 – 4q32.3 site, also appeared to be late-replicating in both cell types (Figure 4-35). In unperturbed RPE1 cells, the site was replicated by forks converging from origins firing in early/mid S-phase, located 0.5 Mb away. In the presence of aphidicolin, the origins surrounding the fragile site appeared to fire slightly earlier, however the fragile location appeared to replicate later. It is possible that an additional origin cluster was fired at this site in late – S phase, however the signal may also represent a termination zone of two forks. The landscape surrounding the region was remarkably similar in the HCT116 cell line, where the site is not fragile. In this cell line origins surrounding the region also appeared to fire earlier in the presence of aphidicolin, but there was not a delay across the fragile core of the site, as observed in the RPE1 cell line.

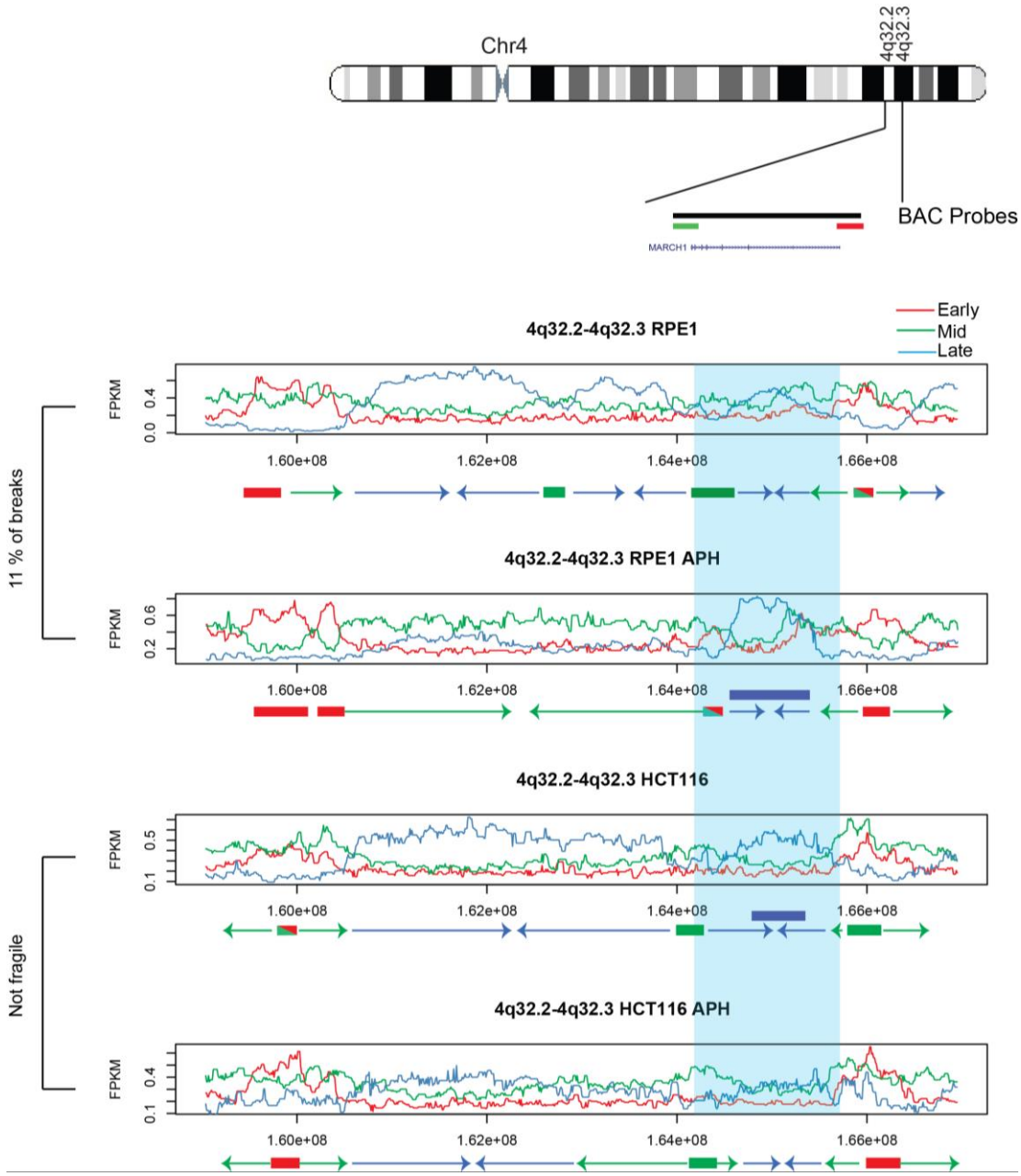


Figure 4-35 Replication landscape at the 4q32.2 -4q32.3 site. FPKM read density in 1000 bp windows are presented for each early (red), mid (green) and late (blue) fractions across FRA1C for the two cell lines in unperturbed conditions and upon aphidicolin treatment. The fragile region is shaded in blue and the locations of the BAC probes used for fine-mapping) are shown at the top, along with the MARCH1 gene. Diagrams of replication dynamics are drawn underneath the graphs, with suspected initiation zones represented as rectangles and travelling forks represented as arrows. At locations where late origin firing cannot be differentiated from termination zones, both arrow and rectangles are drawn.

4.5.1.2 Replication timing across fragile locations in the HCT116 cell line

In HCT116 cells, FRA3B was the most fragile location. Investigating the replication landscape around this site showed that FRA3B span a late replicating zone in both HCT116 and RPE1 cells (Figure 4-36). However, unlike the active sites in RPE1 cells, a clear difference could be seen between the replication profiles in the two cell types across the FRA3B core. In RPE1 cells, where the site is not fragile, FRA3B appeared as a transition zone replicated by forks travelling from early and mid/early initiation zones and converging over the core of FRA3B in late-S. In HCT116, the replication profile was more complex: while similar initiation zones surrounded FRA3B, it appeared that the fragile region spans some late-firing origins, with an extra origin appearing to fire following treatment with aphidicolin.

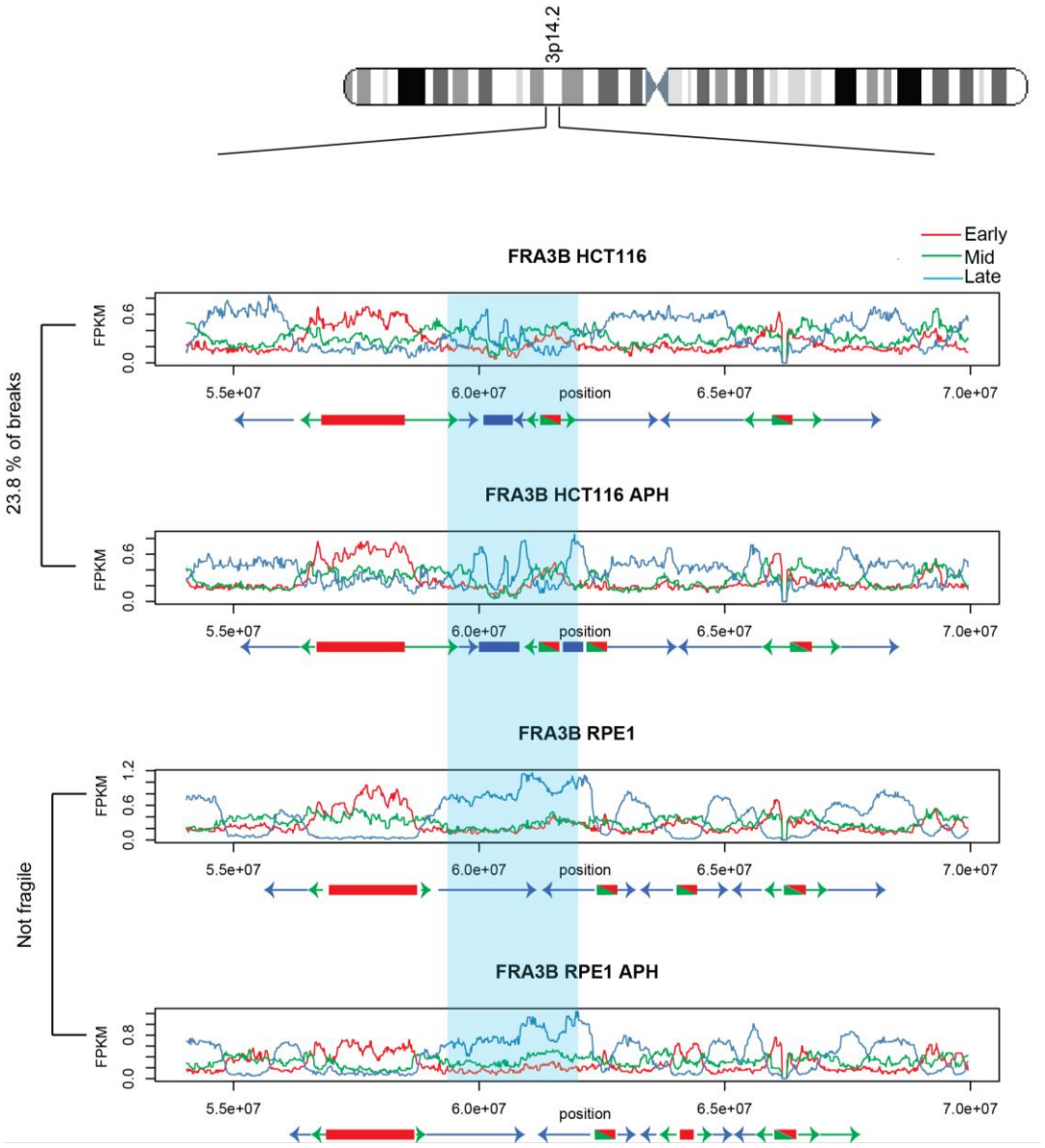


Figure 4-36 Replication landscape at the FRA3B site. FPKM read density in 1000 bp windows are presented for each early (red), mid (green) and late (blue) fractions across FRA3B for the two cell lines in unperturbed conditions and upon aphidicolin treatment. The fragile region is shaded in blue. Diagrams of replication dynamics are drawn underneath the graphs, with suspected initiation zones represented as rectangles and travelling forks represented as arrows. A clear difference in the replication timing profile can be seen between RPE1 and HCT116 cells over the FRA3B fragile region.

FRA4F was the largest CFS region identified during the dine-mapping process: fragility at that site extended over a 10Mb region in the HCT116 cell line. This large region spans a range of replication timing domains, which showed some consistency between the two cell lines (Figure 4-37). In both the RPE1 and HCT116 cells, the centromeric side of FRA4F showed an early to mid-replication timing. Close to the fragile core of the site where most breaks occurred, the profile transitioned towards a late replication timing. In RPE1 cells, where the site is not fragile the transition was relatively smooth, suggesting the region was replicated by forks moving from the early regions. In contrast, sharper peaks were seen in the same region in the HCT116 cells, indicating firing of late origins at the CFS region. In both cell types, the replication profiles shifted to a slightly earlier timing upon aphidicolin treatment.

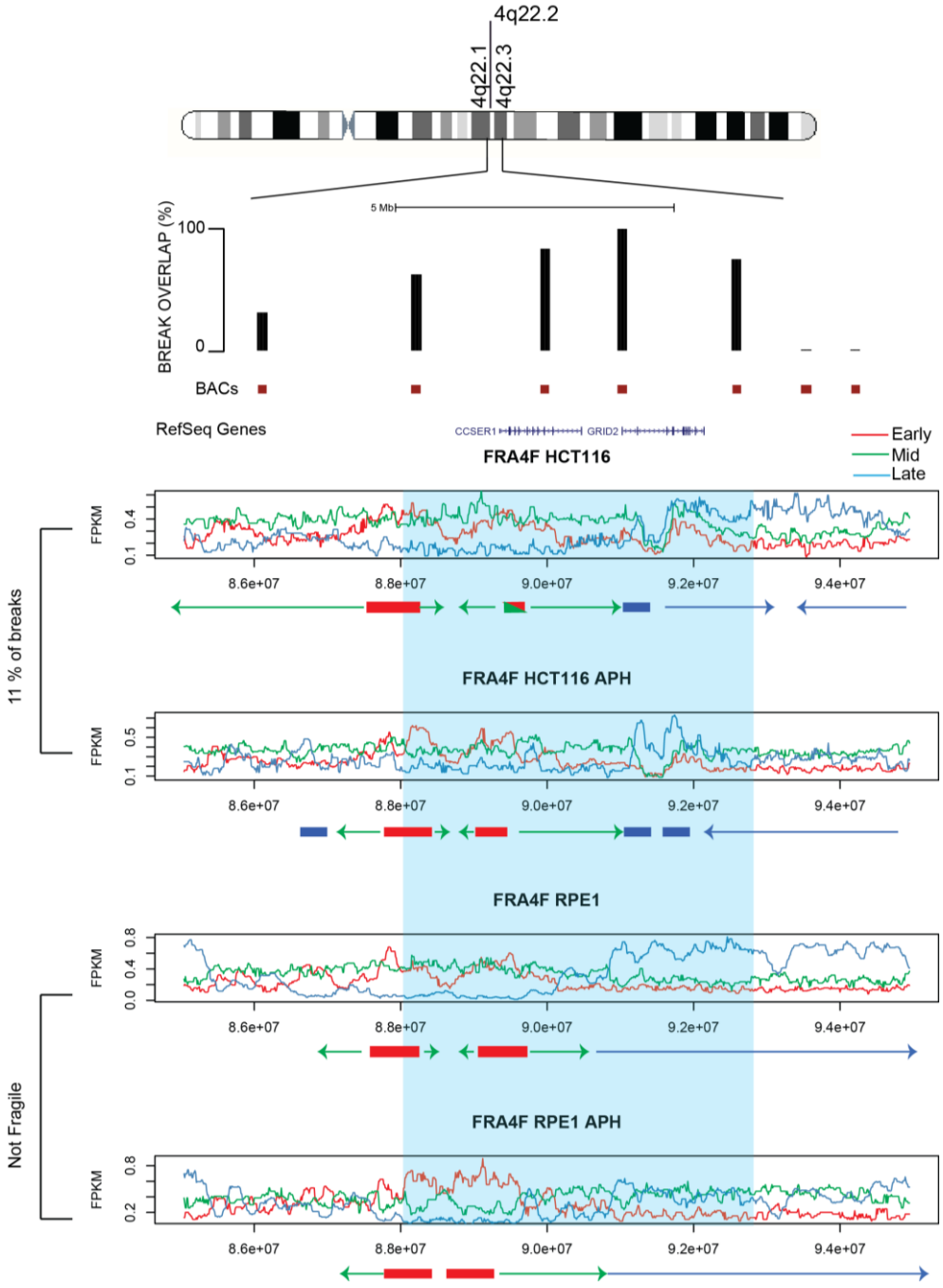


Figure 4-37 Replication landscape at the FRA4F site. FPKM read density in 1000 bp windows are presented for each early (red), mid (green) and late (blue) fractions across FRA4F for the two cell lines in unperturbed conditions and upon aphidicolin treatment. The fragile region is shaded in blue, and the positions of BAC probes used for fine-mapping and their overlap with CFS breaks is shown on the top. Diagrams of replication dynamics are drawn underneath the graphs, with suspected initiation zones represented as rectangles and travelling forks represented as arrows.

Finally, I explored the replication landscape across the FRA2F site at 2q22.2 /2q22.3 (Figure 4-38). This site showed some fragility in both RPE1 and HCT116 cells. Fine-mapping showed that breaks were located telomerically from the LRP1B gene, which is 1.95 Mb long. The region was predominantly late replicating in both cell lines. Strangely, in HCT116 cells, an under-replicated region could be observed in both the control and aphidicolin treated sample, coinciding with the LRP1B gene body, and located centromerically from the break locations identified during fine – mapping. In RPE1 cells, the region showed a mixture between mid and late replication timing. Aphidicolin treatment appeared to increase the signal from mid-reads across the region.

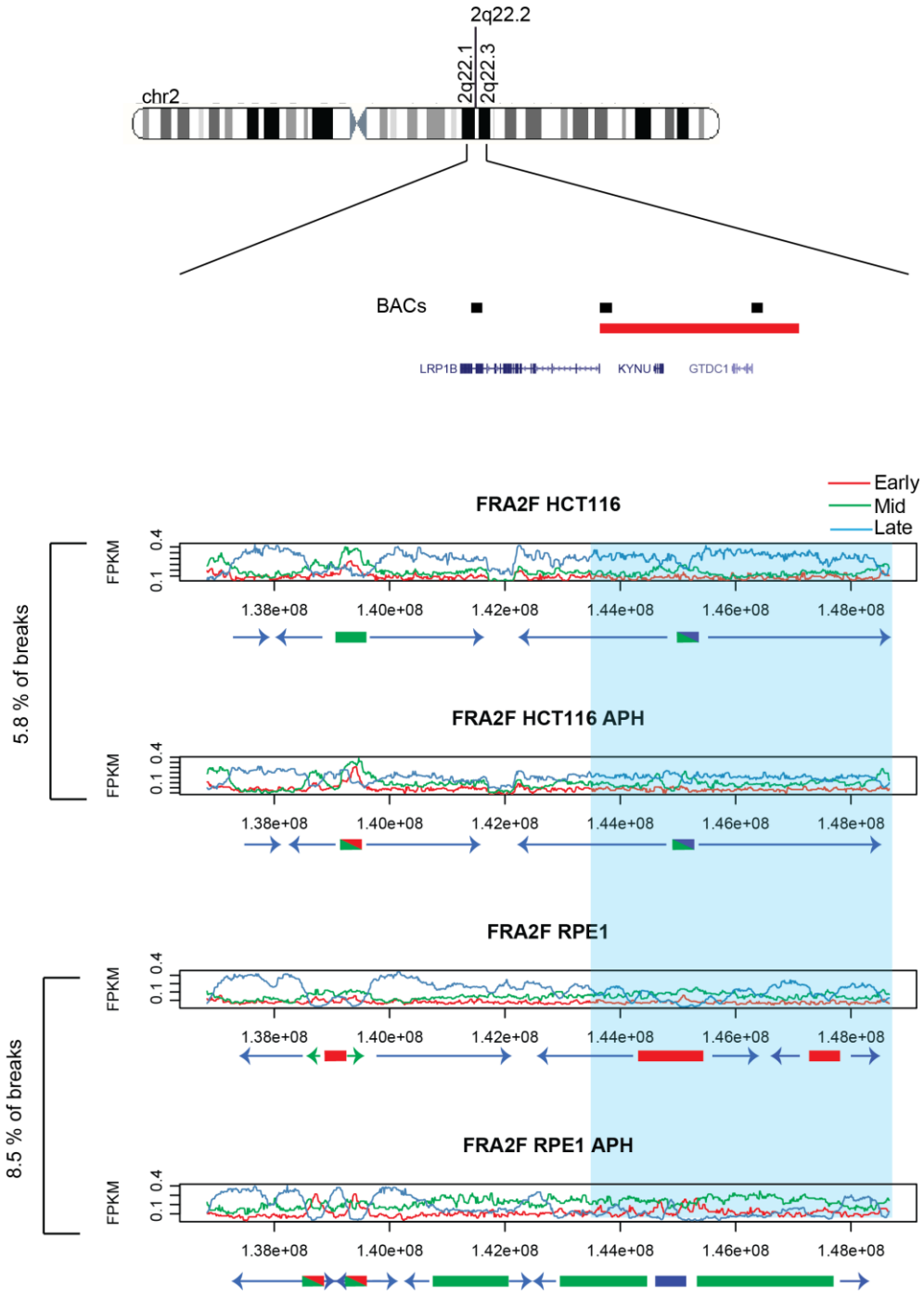


Figure 4-38 eplication landscape at the FRA2F site. FPKM read density in 1000 bp windows are presented for each early (red), mid (green) and late (blue) fractions across FRA4F for the two cell lines in unperturbed conditions and upon aphidicolin treatment. The fragile region is shaded in blue, and the positions of BAC probes used for fine-mapping and their overlap with CFS breaks is shown on the top. Diagrams of replication dynamics are drawn underneath the graphs, with suspected initiation zones represented as rectangles and travelling forks represented as arrows.

4.5.2 CFS regions do not show extreme replication timing changes upon APH treatment

Careful characterisation of the replication timing programme across a number of CFS loci showed that there were no clear and defining features associated with fragility. Since CFS regions behave in a unique manner compared to the rest of the genome upon aphidicolin treatment, I went on to explore whether active CFS are the regions showing most extreme replication timing changes in the presence of replication stress. To do that, I used the Rvalues calculated in 1000bp windows across different conditions. For each 1000bp windows, I subtracted the Rvalue for the aphidicolin treated sample from the Rvalue for the same region under unperturbed conditions. This resulted in positive values for regions which showed a replication delay in the presence of aphidicolin and negative values for regions which replicated earlier upon treatment. I plotted this value, which I called deltaRT, across the genome and investigated how it compared at CFS regions and the rest of the genome.

I first examined the FRA1C site; a visual assessment of the deltaRT across the site showed that there were some changes in replication timing, but they did not appear extreme when compared to the rest of chromosome 1p (Figure 4-39). In HCT116 cells, a transition to later replication timing occurred across most of the fragile region, while in RPE1s, where the site is very unstable, no big changes could be observed. A small spike of negative Rvalues, indicating a transition to earlier replication timing, occurred at the centromeric side of FRA1C. However, similar spikes occurred at other locations throughout chromosome 1p, which were not associated with fragile regions. Therefore, a visual inspection of the differential timing did not reveal a huge response to APH treatment at FRA1C.

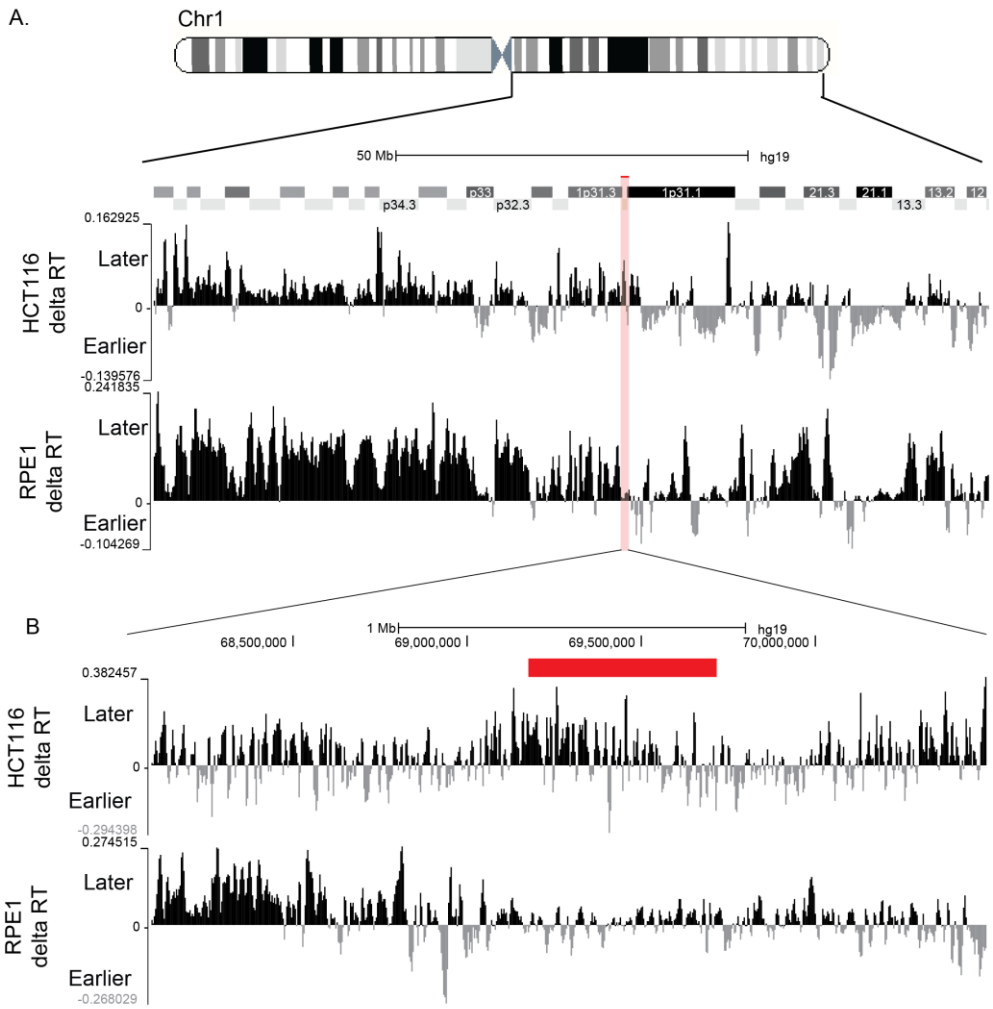


Figure 4-39 Replication timing changes across chromosome 1p and FRA1C in the presence of aphidicolin. deltaRT values were calculated in 1000 bp windows and plotted across chromosome 1p (A) and FRA1C (B) for both the RPE1 and the HCT116 cell line to determine if extreme replication timing changes are seen at FRA1C compared to the rest of chromosome 1. A. delta RT values are plotted across chromosome 1p for the HCT116 cell line (top) and the RPE1 cell line (bottom). Positive deltaRT values indicate later replication timing in the presence of aphidicolin and negative deltaRT values indicate earlier replication upon aphidicolin treatment. The location of FRA1C is indicated in red. No extreme replication timing shifts are seen across FRA1C. B. delta RT values across the FRA1C region in HCT116 cells (top) and RPE1 cells (bottom). The fragile core of the region, as defined by FISH fine-mapping, is indicated by a red bar. A slight delay across the region is seen in the HCT116 cell line, while no change can be observed in the RPE1 cell line, where the region is fragile.

To examine the changes in more detail, I plotted the frequency distribution of delta RT values across FRA1C region and compared it to the distribution of values across all of chromosome 1 (Figure 4-40). In RPE1 cells, I found that the distribution of RT values was slightly shifted towards a later replication timing upon aphidicolin treatment, compared to the distribution for all of chromosome 1. Surprisingly, this shift was a lot more pronounced in HCT116 cells, where the region was not as strongly fragile. This observation strongly suggests that the small replication delay seen across the site for RPE1 cells is not a determinant of instability at the site.

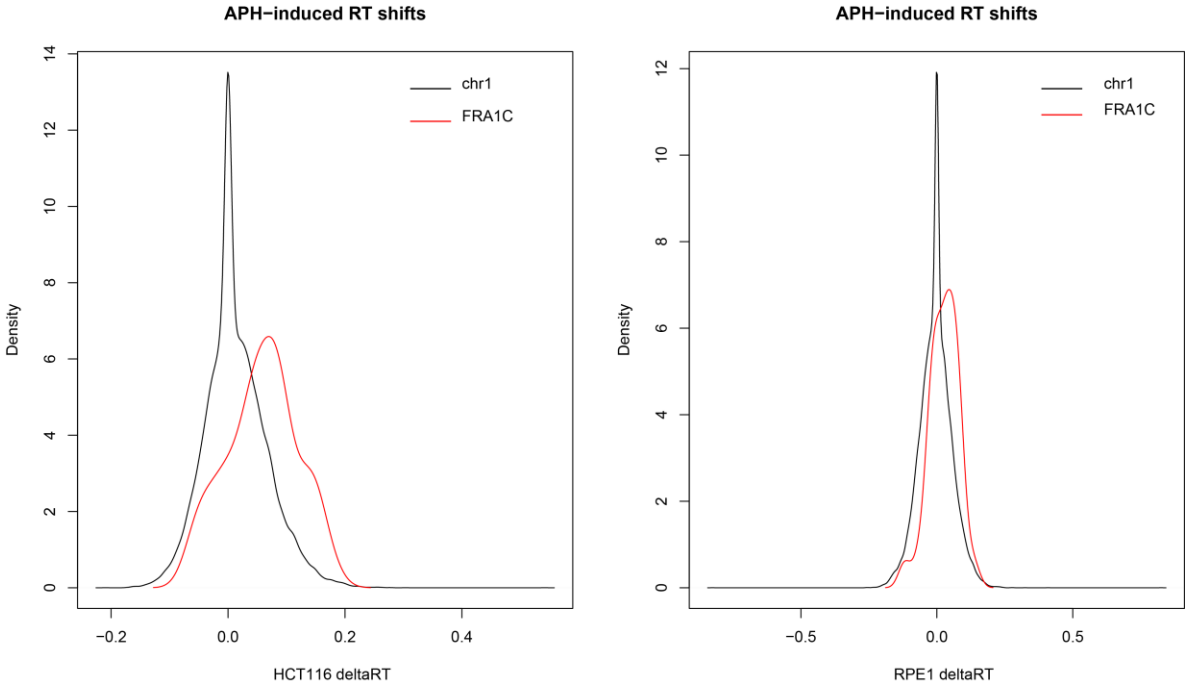


Figure 4-40 Distribution of deltaRT values across FRA1C and chromosome one in the two cell lines. The frequency distribution of deltaRT values was plotted for FRA1C (in red) and all of chromosome 1 (in black) in HCT116 and RPE1 cells. A small shift towards positive values, indicating a replication delay, was observed in the RPE1 cell type. The shift was even more pronounced in the HCT116 cell type, where breaks are formed at the site less frequently than in RPE1 cells.

Finally, I also compared the relationship between Rvalues in the control and the aphidicolin-treated samples across FRA1C and chromosome 1. I plotted the R value for each 1000 bp window across chromosome 1 in control condition versus the R value in the same window under aphidicolin treatment (Figure 4-41). A relatively good correspondence could be seen between the two values for most regions.

However, some windows showed a discordance between the two values and appeared as outliers, displaying either lower or higher Rvalues in the aphidicolin treated sample compared to the control. Windows within the FRA1C site were not among the outliers and instead, appeared to be similarly affected by aphidicolin to the rest of chromosome 1.

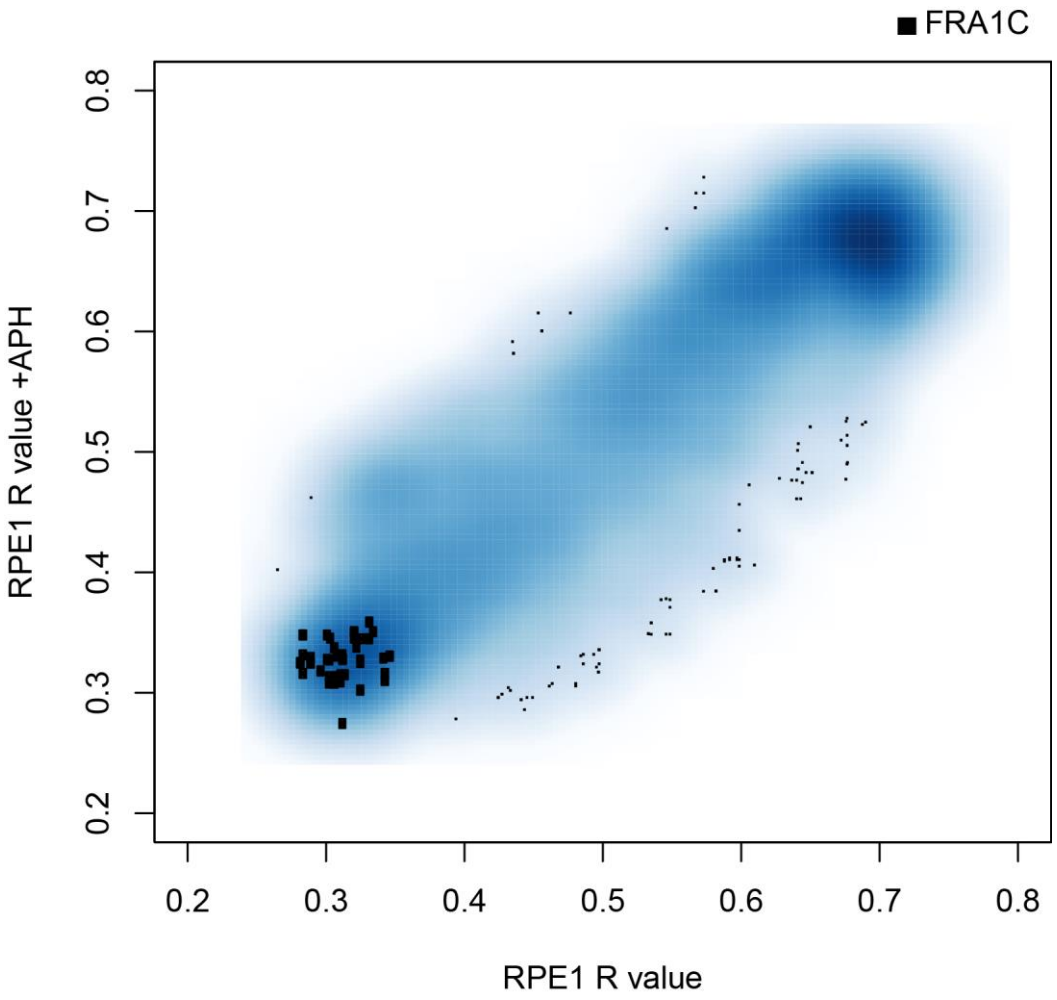


Figure 4-41 Relationship between Rvalues in control conditions and upon aphidicolin treatment on chromosome 1 and at the FRA1C locus. Rvalues were plotted for the control sample (on the x-axis) and the aphidicolin-treated sample (on the y-axis). Rvalues across chromosome 1 are shown as a blue scatter, while values for FRA1C are shown as black rectangles.

I next assessed the changes at the novel fragile site in RPE1 cells, located at the 4q32.2-4q32.3 boundary (Figure 4-42). This site behaved very differently to FRA1C: a region of positive delta RT values, indicating a replication delay in the presence of

aphidicolin, could be seen across the site. Furthermore, the delay was specific to RPE1 cells, where the site is fragile; in HCT116s, the opposite trend for earlier replication timing was observed. Similarly to the FRA1C region, comparison of the delta RT values at this site with a larger surrounding region in the chromosome 4q arm revealed that the delay at this site was not the most extreme within the region: locations with larger delta RT values were seen, which did not correspond to active fragile locations.

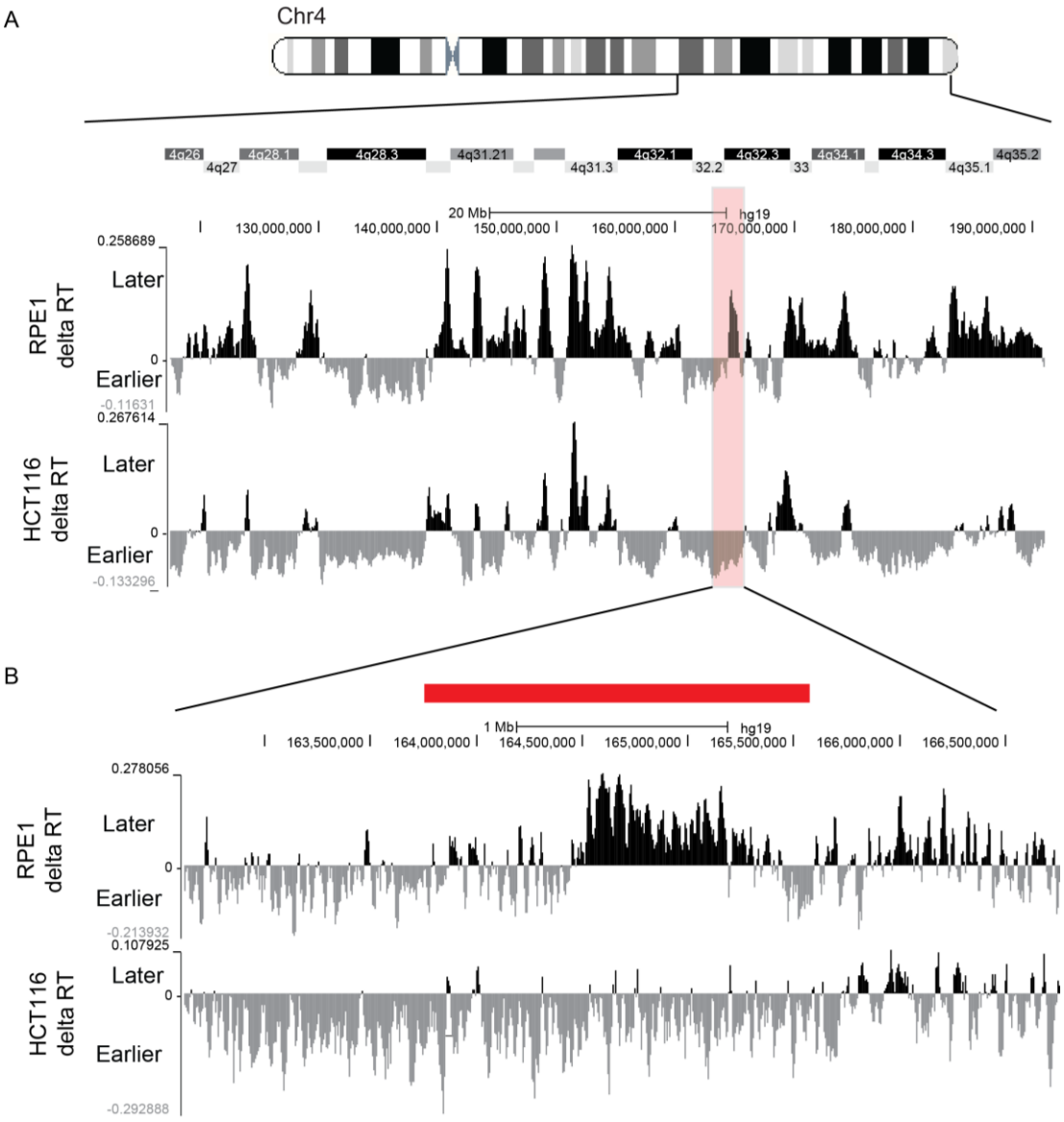


Figure 4-42 Replication timing changes across chromosome 4q and the 4q32.2/4q32.3 site in the presence of aphidicolin. deltaRT values were calculated in 1000 bp windows and plotted across a 20Mb region on chromosome 4q (A) and the 4q32.2-4q32.3 region(B) for both the RPE1 and the HCT116 cell line to determine if extreme replication timing changes are seen at the fragile site compared to the surrounding region. A. delta RT values are plotted across chromosome 4q for the RPE1 cell line (top) and the HCT116 cell line (bottom). Positive deltaRT values indicate later replication timing in the presence of aphidicolin and negative deltaRT values indicate earlier replication upon aphidicolin treatment. The location of the CFS is indicated in red: no extreme replication timing shifts are seen across the fragile region B. delta RT values across the 4q32.2-4q32.3 CFS region in the two cell lines. The fragile core of the region is indicated by a red bar.

When I compared the frequency distribution of delta RT values across the fragile region to the distribution across the whole chromosome 4 I observed a small but clear shift to higher values, indicating this region experienced a replication delay compared to the rest of the chromosome (Figure 4-43). This was also confirmed by a comparison of the Rvalues under control conditions and in the presence of aphidicolin across the chromosome and at the fragile region: windows within the fragile region showed a bigger change in Rvalues compared to the rest of the genome. This confirms that fragility at this site is accompanied by a cell-type specific delay in replication timing upon APH treatment.

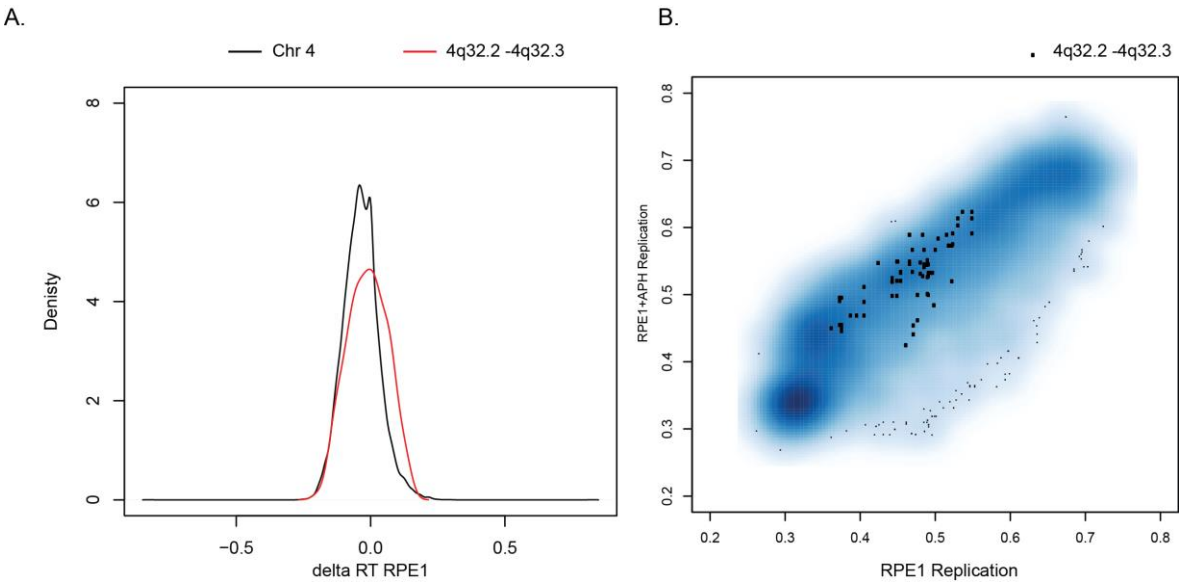


Figure 4-43 Replication timing changes at the 4q32.2-4q32.3 fragile location in the RPE1 cells. A. Distribution of deltaRT values across the CFS and chromosome 4. The frequency distribution of deltaRT values is shown for the CFSC (red) and all of chromosome 4 (black) in RPE1 cells. A shift towards higher values, indicating later replication timing can be observed for the fragile site. B. Relationship between Rvalues in control conditions and upon aphidicolin treatment on chromosome 4 and at the CFS region. Rvalues for the control sample are shown on the x-axis and Rvalues for the aphidicolin-treated sample are on the y-axis. Rvalues across chromosome 4 are shown as a blue scatter, while values for the CFS region are shown as black rectangles.

Next, I examined CFS sensitivity to aphidicolin in the HCT116 cell line.

FRA3B, the most fragile location in this cell type, displayed changes in both directions: the core of the site spanned changes from towards an earlier and a later

replication timing. None of the changes appeared extreme when compared to the landscape across the chromosome 3p arm (Figure 4-44). Less changes were seen across the CFS region in the RPE1 cell line, where FRA3B is not fragile. Comparison of the distribution of delta RT values for the FRA3B region to all of chromosome 3 revealed a very similar distribution between the two, indicating that FRA3B does not show extreme changes in response to aphidicolin (Figure 4-45). A comparison of R values in control conditions and following aphidicolin treatment showed that windows across FRA3B deviated slightly in their response compared to the general trend for chromosome 3. Specifically, a number of windows appeared to shift to an earlier replication timing upon replication stress induction.

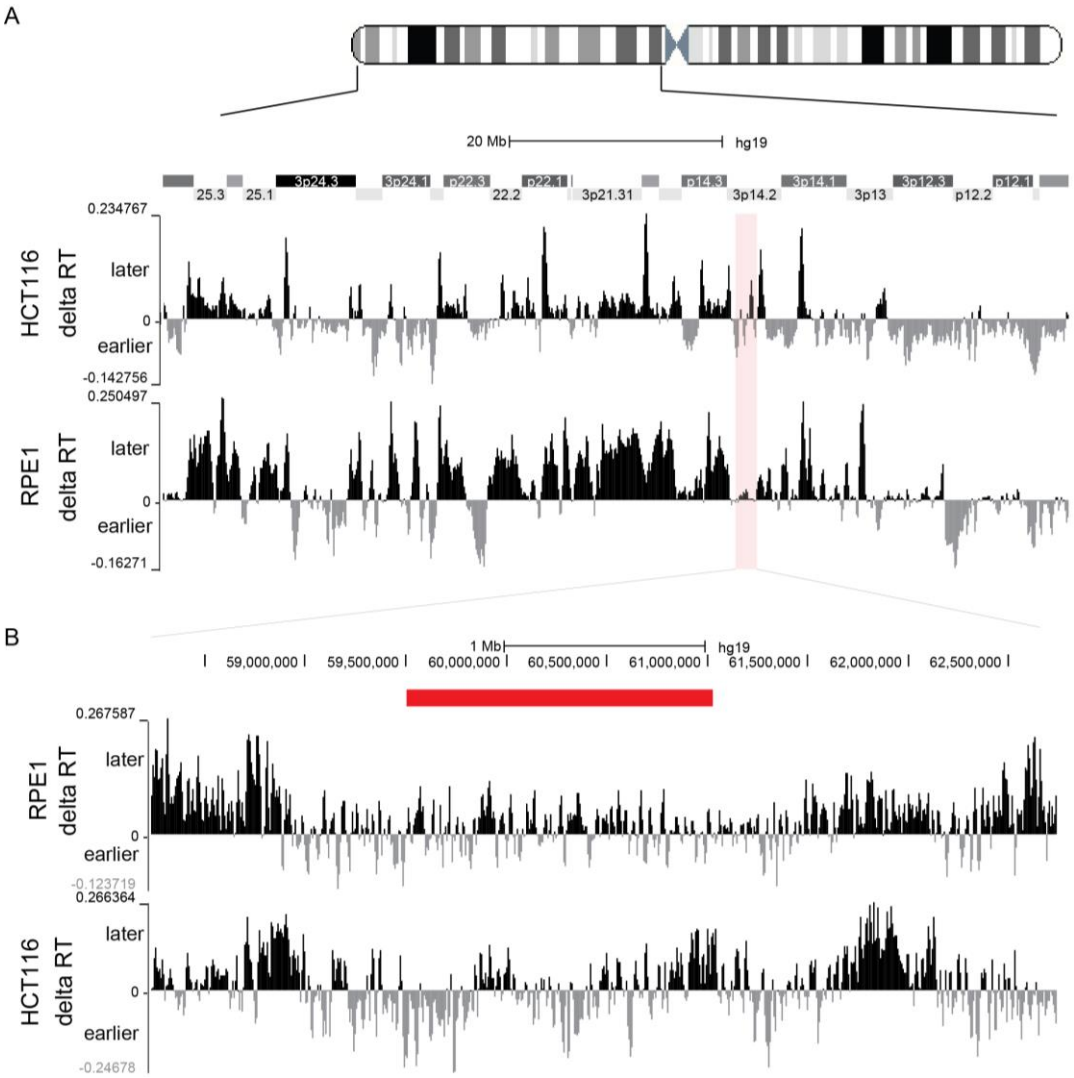


Figure 4-44 Replication timing changes across chromosome 3p and the FRA3B site in the presence of aphidicolin. deltaRT values were calculated in 1000 bp windows and plotted across chromosome 3p (A) and the FRA3B region(B) for both the RPE1 and the HCT116 cell line. A. delta RT values are plotted across chromosome 3p for the HCT116 cell line (top) and the RPE1 cell line (bottom). The location of the CFS is indicated in red: no extreme replication timing shifts are seen across the fragile region B. delta RT values across the FRA3B CFS region in the two cell lines. The fragile core of the region is indicated by a red bar.

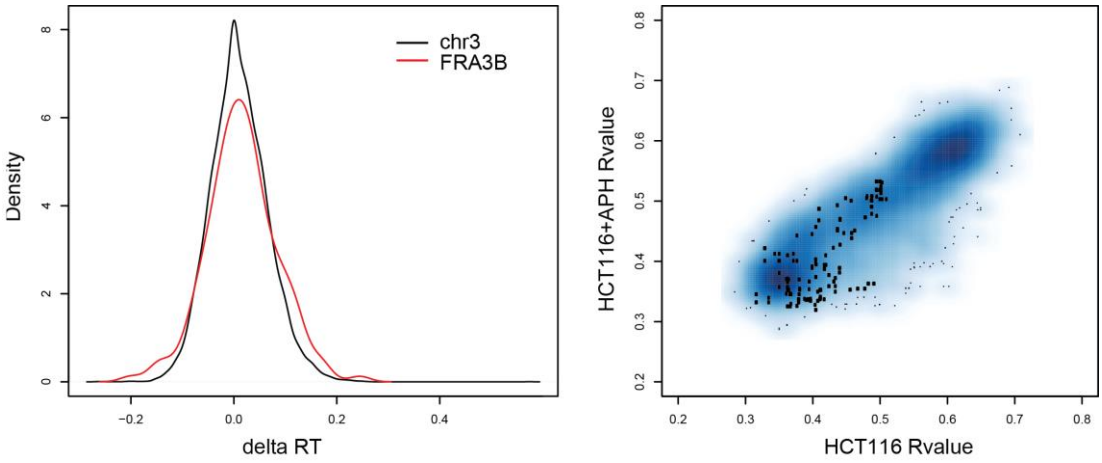


Figure 4-44 Replication timing changes at the FRA3B fragile location in the HCT116 cells. A. Distribution of deltaRT values across the CFS and chromosome 3. The frequency distribution of deltaRT values is shown for the CFS (in red) and all of chromosome 3 (in black) in HCT116 cells. B. Relationship between Rvalues in control conditions and upon aphidicolin treatment on chromosome 3 and at the FRA3B region. Rvalues for the control sample are shown on the x-axis and Rvalues for the aphidicolin-treated sample are on the y-axis. Rvalues across chromosome 3 are shown as a blue scatter, while values for the CFS region are shown as black rectangles.

Finally, I examined FRA4F, a site responsible for 11% of breaks in the HCT116 cell line, which was not fragile in RPE1 cells. Surprisingly, this site changed to an earlier replication timing upon treatment with aphidicolin (Figure 4-45). This change appeared to be specific to the HCT116 cell line, where the site is fragile and did not occur in RPE1s. Similarly to the other CFS regions analysed, it appeared that the change this fragile site was not larger than the changes seen across chromosome four (Figure 4-46). Consistently, windows from that site did not appear as outliers when Rvalues from the control sample were plotted versus the Rvalues from the aphidicolin treated sample.

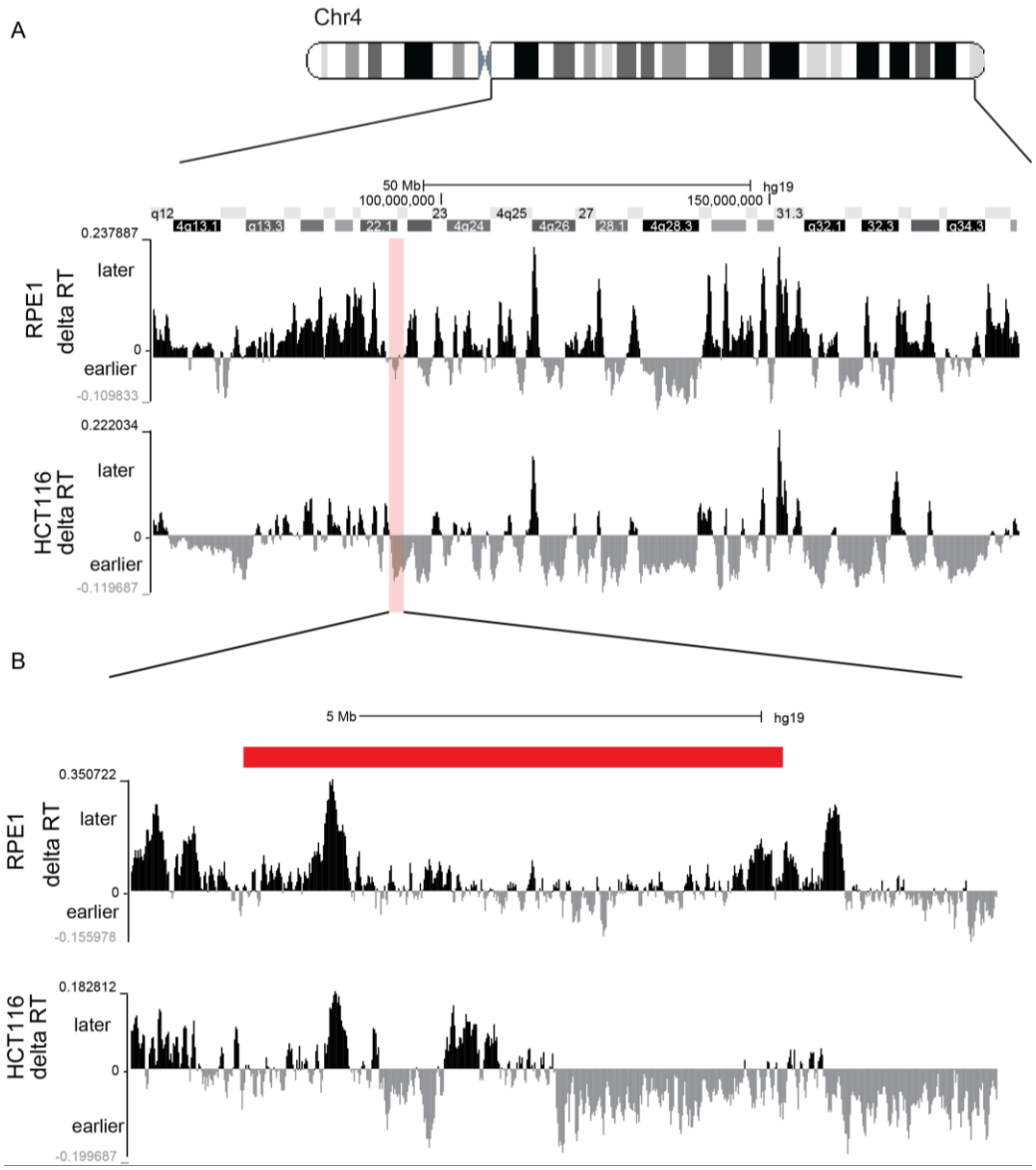


Figure 4-45 Replication timing changes across chromosome 4q and the FRA4F site in the presence of aphidicolin. deltaRT values were calculated in 1000 bp windows and plotted across chromosome 4q (A) and the FRA4F region(B) for both the RPE1 and the HCT116 cell line. The location of the CFS is indicated in red: no extreme replication timing shifts are seen across the fragile region.

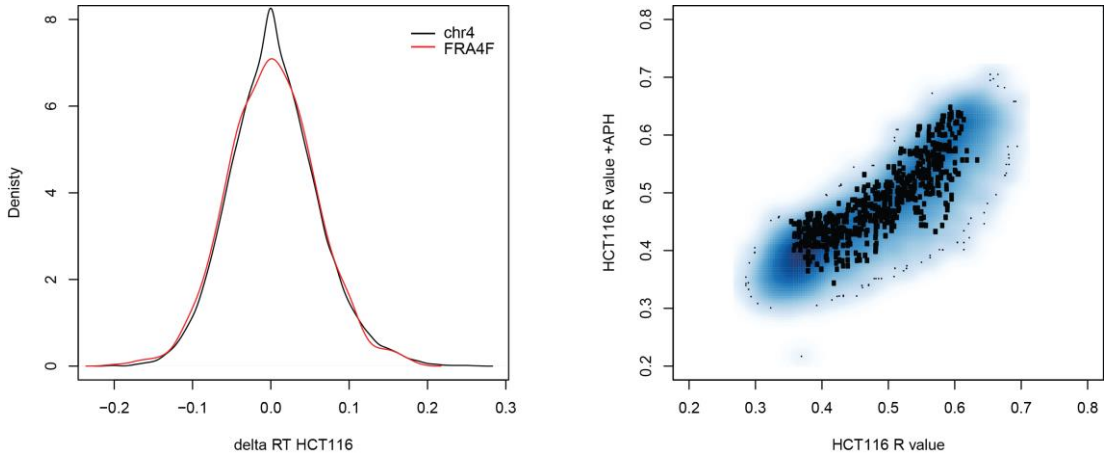


Figure 4-46 Replication timing changes at the FRA4F fragile location in the HCT116 cells. A. Distribution of deltaRT values across the CFS and chromosome 3. The frequency distribution of deltaRT values is shown for the CFS (red) and all of chromosome 4 (black) in HCT116 cells. B. Relationship between Rvalues in control conditions and upon aphidicolin treatment on chromosome 4 and at the FRA4F region. Rvalues for the control sample are shown on the x-axis and Rvalues for the aphidicolin-treated sample are on the y-axis. Rvalues across chromosome 4 are shown as a blue scatter, while values for the CFS region are shown as black rectangles.

In summary, CFS regions do not appear to be uniquely sensitive to aphidicolin. While some sites display changes in their replication timing, the changes induced vary in their directionality and do not appear to be more extreme than the rest of the genome. Therefore, my Click-seq data does not support a model for CFS fragility based on a simple replication timing delay in the presence of aphidicolin.

4.5.3 Regions of the genome showing most extreme replication timing changes in the presence of APH

As CFS regions did not show extreme sensitivity to aphidicolin, I set out to determine which genomic locations show the largest changes in replication timing upon aphidicolin treatment. I examined the distribution of deltaRT values across the genome and selected windows that had deltaRT values removed at least three standard deviations from the mean. Surprisingly, I found that some of these clustered in similar locations for both the RPE1 and the HCT116 cell lines. It

appeared that these clusters of extreme changes were formed around large domains of early or late replication timing and probably corresponded to disruption of domain boundaries observed upon APH treatment.

4.6 Discussion

In this chapter, I have described the optimisation and application of a novel variant of the replication timing mapping method Repli-seq, which I have termed Click-seq. While I have not performed a direct comparison between the Repli-seq and the Click-seq methodology and the data generated using the two methods, Click-seq appears to clearly delineate early and late replicating regions of the genome with good reproducibility. The Click-seq protocol is faster than the Repli-seq protocol and the two-hour BrdU labelling pulse recommended in Repli-seq can be replaced by a shorter, 30 minute EdU pulse, allowing better resolution of replication timing domains. An unresolved issue of the Click-seq protocol is the generation of reads that cannot be matched to the human genome reference sequence. However, this appeared to be connected to the quality and the concentration of the sequencing libraries, suggesting that consistent preparation of high-quality libraries would reduce the problem.

Utilising the Click-seq methodology, I assessed genome-wide replication timing patterns in the RPE1 and HCT116 cell lines; although the replication programmes of many different cell types have been analysed via Repli-seq and Repli-chip as a part of the ENCODE project, the two cell types used in this project had not been studied yet. The general characteristics of the replication programme within these two cell types are consistent with the well-established rules of replication timing: early, GC-rich regions containing transcriptionally active genes replicate earlier, while gene-poor, GC-poor regions tend to have a later replication timing. The RPE1 cell type, which derives from normal tissue, appeared to show more defined replication timing profiles, indicating less variation at the individual cell level. Domain size in this cell type was also smaller and the domains appeared less contiguous than in the tumour-derived HCT116 cells, which again may be due to more controlled

progression of replication. In addition to studying the replication timing profiles of the two cell types under unperturbed conditions, I also investigated the effect of replication stress on the replication programme of the two cell types. Surprisingly, rather than a universal delay across the genome, I found that aphidicolin induced bi-directional locus-specific changes in replication timing, with some regions replicating earlier and some regions replicating later. In addition, the low concentration of aphidicolin used in this study caused a subtle loss of features normally associated with early or late replication timing: early replicating regions in the aphidicolin-treated sample showed lower GC content and a lower density of expressed genes compared to early regions in control sample, while the inverse was true for late regions. Therefore, I can conclude that pharmacologically induced replication stress causes a mis-regulation of replication timing, rather than a genome-wide shift towards later replication. Recently, a similar observation was made about *Rif1*, a genome-wide regulator of replication timing: depletion of the protein resulted in bi-directional shifts in replication timing. However, the shifts observed upon *Rif1* depletion were more significant than the aphidicolin-induced changes (Foti et al. 2016). As aphidicolin can mimic endogenous replication stress present in CIN+ve colorectal cancer cell line, it is tempting to speculate whether similar shifts in replication timing can be observed in tumour cells (Burrell et al. 2013). A shift towards later to earlier replication timing could explain some features of cancer cells, such as abnormal DNA methylation.

Surprisingly, investigation of the replication timing landscape across active fragile sites within the two cell types in the presence and absence of aphidicolin failed to reveal e signature features of replication associated with CFS expression. As expected, most CFS regions span late replicating regions; two of the most fragile CFS locations within the RPE1 cell line showed remarkably similar replication patterns, composed of a late replication zone, replicated by long-travelling forks initiating from origin zones with an earlier replication timing. Surprisingly, aphidicolin treatment caused a shift towards earlier replication timing at the initiation zones surrounding the fragile regions at both FRA1C and 4q32.2-4q32.3

sites. The replication profiles of highly fragile locations within the HCT116 cell line showed less similarity; a common feature of FRA3B and FRA4F was the appearance of peaks in the late-replication track not present in RPE1 cells, suggesting possible firing of late origins both in control cells and in the presence of aphidicolin. Finally, a comparison of the replication timing of these sites in control and aphidicolin-treated cells indicated that their fragility is not rooted in an extreme change in replication timing in the presence of aphidicolin, but is likely a result of more subtle characteristics of the replication dynamics of these sites.

5 Chapter 5: Replication stress and interphase chromatin state at CFS

Fragility at CFS regions extends over large genomic distances: the smallest site identified in this study, FRA1C, spans 0.5 Mb, while breaks at FRA4F extend over a 5 Mb genomic segment. Historically, CFS regions have been defined as affecting an entire chromosome band, or a boundary between two cytogenetic bands, encompassing millions of base pairs of genomic sequence (Durkin & Glover 2007; Debatisse et al. 2006). These observations strongly suggest that fragility at CFS is a feature associated with large-scale chromatin structures, at the level of organisation well above the 30nm fibre. However, no studies to date have explored the chromatin dynamics at CFS regions: in fact, only limited efforts have been made to study chromatin at CFS regions, with a focus on chromatin composition rather than structure. A 2009 study found that the most unstable CFS regions in lymphoblastoid cell lines were hypo-acetylated compared to the regions surrounding them and treatment with the deacetylase inhibitor TSA decreased the frequency of breaks (Jiang et al. 2009); the H3K4me1 mark has also been identified as a significant predictor of CFS fragility in a computational analysis which did not take into account the cell type specificity of breakage (Fungtammasan et al. 2012).

Chromatin structure is the background for all of the processes implicated in CFS fragility and has never been assessed as a potential contributor. In this chapter, I investigate large-scale chromatin structure at the FRA1C and FRA3B fragile regions using fosmid pair FISH. This approach is based on two fosmid pairs, mapping to the region of interest and separated by a linear distance ranging between 100 kb and 1.5 Mb, which are hybridised to a cell population; the distance between the two fosmids signals is then measured in a large number of nuclei. The distribution of distances across the population is reflective of the underlying large-scale chromatin

states. This approach has been used successfully to demonstrate altered large scale structure at the inactive X chromosome and during development (Naughton et al. 2010; Williamson et al. 2012). Employing this approach, I characterise the effect of replication stress induction on chromatin surrounding the sites in both asynchronous populations and synchronised cultures transitioning through S-phase. I also compare aphidicolin-induced replication timing changes to alterations in large-scale chromatin structure.

5.1.1 Replication stress effect on chromatin compaction in asynchronous cell populations

I first set out to explore how aphidicolin affects the chromatin landscape across different genomic regions in the RPE1 cell line. As well as the fragile FRA1C region and FRA3B, which is not active in this cell type, I investigated the effects of replication stress on two chromosome 11 loci with differing functional characteristics. This allowed me to differentiate between effects associated with CFS regions from genome-wide effects of replication stress, including direct effects on chromatin structure and indirect effects, such as biases arising from cell cycle differences between the control and the aphidicolin-treated sample. I selected fosmid probe pairs surrounding the different locations and hybridised them to PFA-fixed cells, grown either in unperturbed conditions or in the presence of various concentrations of aphidicolin. The length of aphidicolin treatment was always 24 hours. I measured the distance between the probe pairs in a large number of nuclei and compared how the distribution of distances varied between loci and upon induction of replication stress.

5.1.2 Interphase chromatin compaction at 11p14.1 and 11p15.1

To define if replication stress exerts genome-wide effects on chromatin structure, I first assessed chromatin response to aphidicolin at two chromosome 11 locations with differing functional properties: the gene-rich 11p15.1 and the gene-poor

11p14.1. In addition to gene-density, these two loci are known to differ in their chromatin structure, with 11p15.1 showing negative supercoiling, associated with open chromatin and 11p14.1 forming a domain of positive supercoiling with compact chromatin architecture (Naughton et al. 2013). Previously, a CFS was identified at chromosome 11p15.1 (FRA11C) and the region was found to overlap with breakpoints in chromosomal rearrangements in bladder cancer (Moriarty & Webster 2003). However, no breaks were observed within that region in RPE1 cells. Although the fosmid pairs at the two loci were located a similar distance apart, the physical distances between the fosmids at the 11p14.1 locus were much smaller than at 11p15.1, likely due to the more compact conformation of the positively supercoiled, gene poor region (Figure 5-1). Aphidicolin treatment did not cause a significant change in compaction at either of the two loci. This suggested that replication stress does not exert a genome-wide effect on chromatin structure.

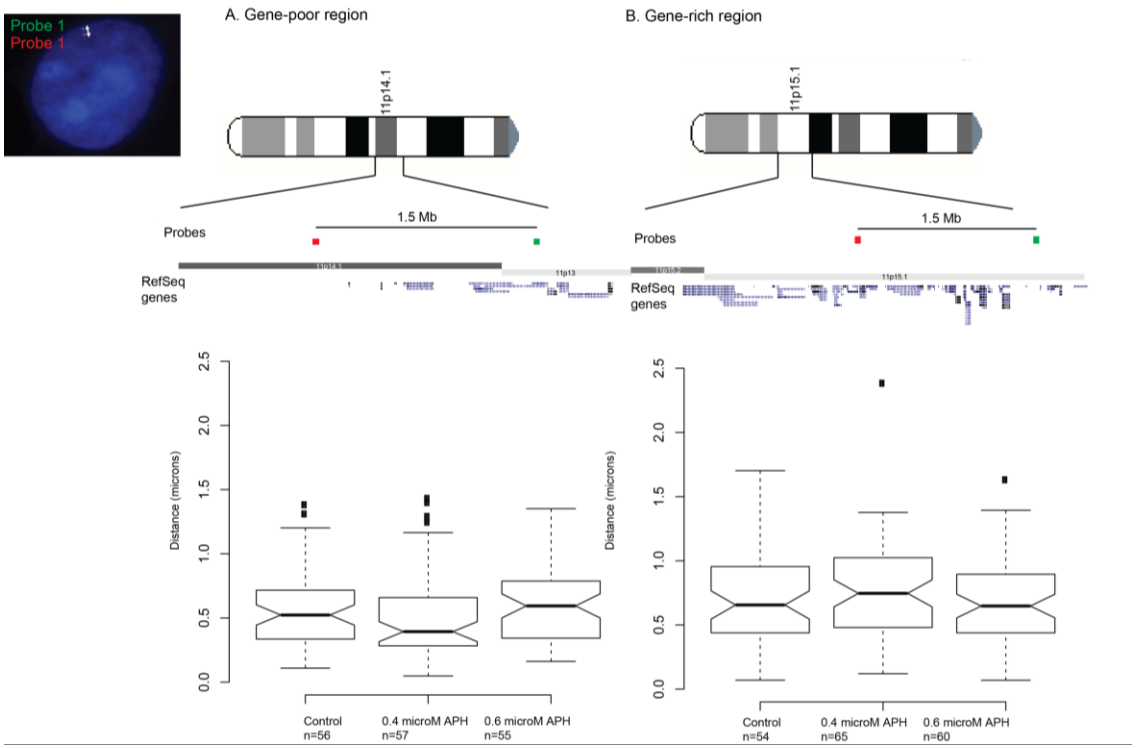


Figure 5-1 Chromatin compaction at the gene-poor 11p14.1 region (A) and the gene-rich 11p15.1 locus (B) in RPE1 cells upon aphidicolin treatment. RPE1 cells were grown under control conditions or in the presence of two different concentrations of aphidicolin. Cells were then hybridised to pairs of fosmid probes hybridising to either 11p14.1 or 11p15.1. Fosmid pairs were located 1.5 Mb apart. Probe positions are shown in red and green in the top diagram. The bottom graphs show a boxplot of the distance between the two fosmids across the different samples. Numbers of nuclei included in the analysis is shown for each category. Inset, a representative image showing a nucleus with hybridised probes labelled in red and green.

5.1.3 Interphase chromatin compaction at FRA1C

I next examined the consequences of aphidicolin treatment at the FRA1C locus, the most fragile region in the RPE1 cell line. Surprisingly, at FRA1C, I found that aphidicolin treatment caused a change in chromatin state towards a more compact conformation (Figure 5-2). Significant compaction of chromatin was observed in cells treated with 0.6μM APH, and the trend for a more compact state was also present in the population treated with just 0.4μM of the drug. As this compaction was not observed at the chromosome 11 loci, I could exclude the possibility that it is

due to a genome-wide effect of aphidicolin. To investigate the chromatin change further, I examined the frequency distribution of distances across the control and the two aphidicolin treated samples; in particular, I wanted to determine if there was evidence of a bi-modal distribution in the presence of aphidicolin, indicating two distinct conformations of the locus, corresponding to a stable and a fragile state. A small extra peak could be observed among the distribution of distances for cells treated with 0.4 μM aphidicolin, but due to a small sample size (n=56), it was difficult to determine if the peak reflected two distinct states of the locus. However, there was no sign that two distinct chromatin states existed at the locus upon treatment with 0.6 μM APH: instead, a shift in frequency towards a more compact state in the treated samples compared to the controls appeared (Figure 5-3). This observation was highly surprising, and counter intuitive to the expectation that a break at the region would result in an increase in inter-fosmid distance, due to the physical separation of the probes. The fact that this observation was made in an asynchronously growing cell population which has been treated with aphidicolin for the length of a full cell cycle made interpretation difficult. Specifically, it was unclear whether compaction occurred prior to the formation of metaphase lesions at the site or as a consequence of the instability at FRA1C. I therefore decided that experiments in synchronised cell populations were necessary to delineate chromatin dynamics at the site and determine if replication stress – associated compaction contributed to lesion formation.

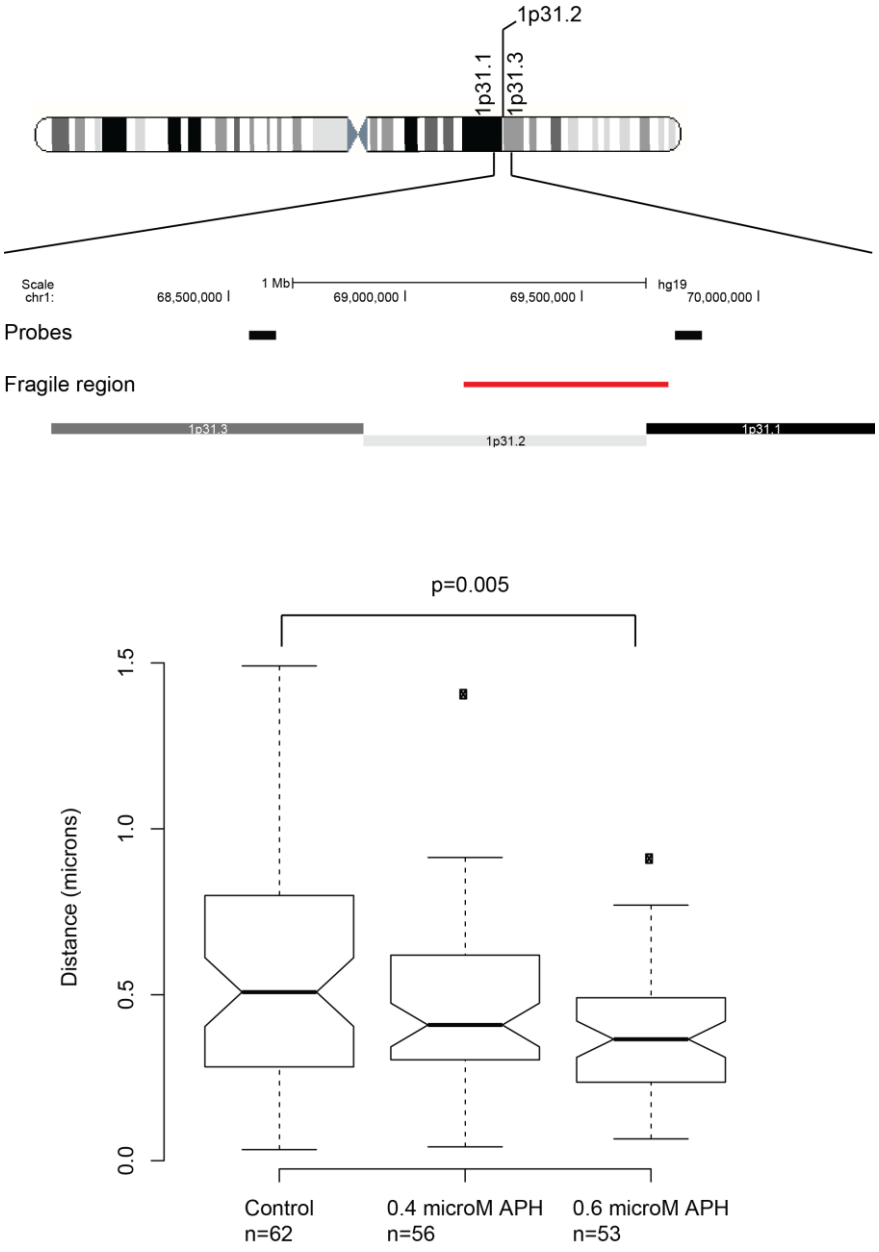


Figure 5-2 Chromatin compaction changes at the FRA1C locus in RPE1 cells upon aphidicolin treatment. RPE1 cells were grown under control conditions or in the presence of two different concentrations of aphidicolin. Cells were then hybridised to a pair of fosmid probes located 1.1 Mb apart, which surrounded the FRA1C locus. Probe positions are shown as black bars in the top diagram; the fragile region is shown as a red bar. The bottom graph shows a boxplot of the distance between the two fosmids across the different samples. Numbers of nuclei included in the analysis is shown for each category. P value is for a Wilcoxon test Data is representative of two biological replicates.

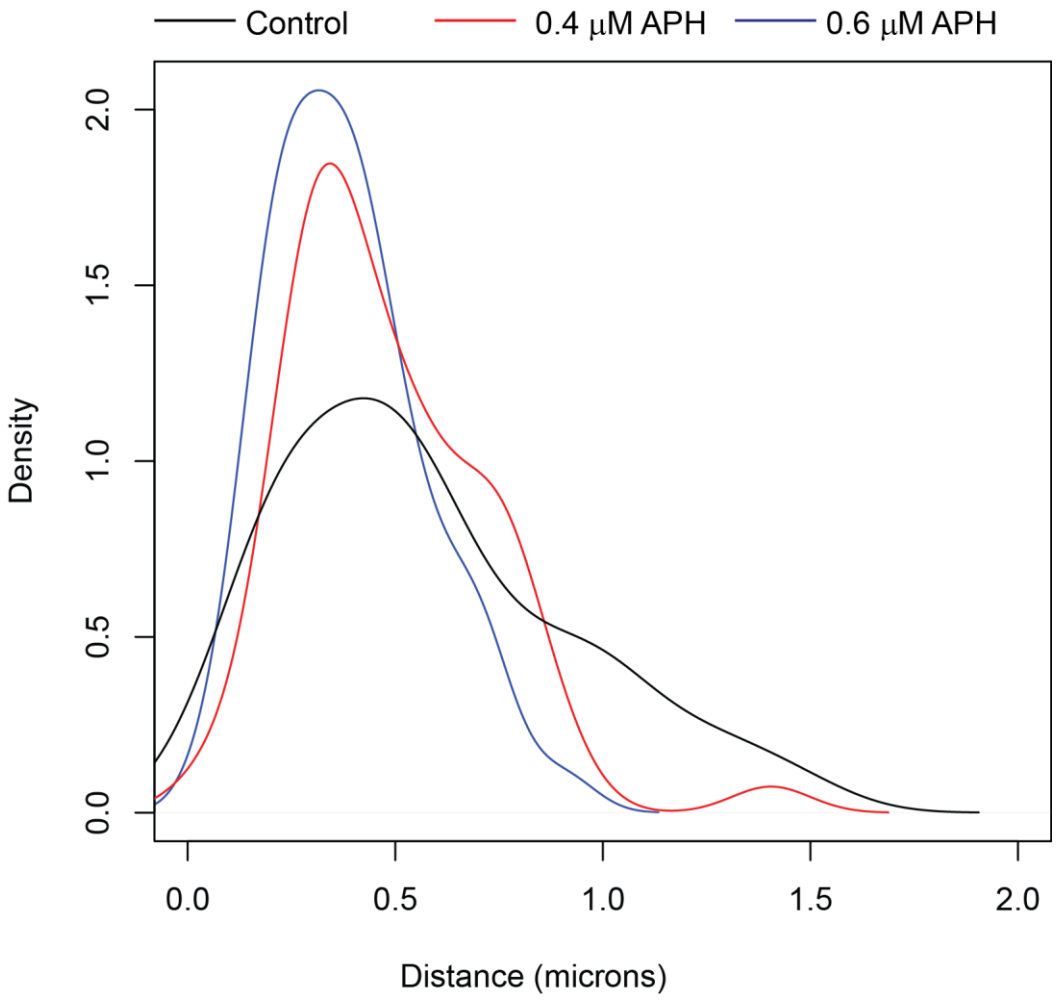


Figure 5-3 Frequency distributions of fosmid pair distances at the FRA1C locus. Distributions are shown for the control sample (black), and 0.4 μM (red) and 0.6μM (blue) aphidicolin treatment.

5.1.4 Interphase chromatin compaction at FRA3B

I next wanted to determine if the effect of aphidicolin observed at FRA1C is specific to this active fragile site or can also be observed at a non-expressed CFS location, such as FRA3B. I selected fosmid probes surrounding the FHIT gene and performed fosmid FISH in unperturbed RPE1 cells and under the same aphidicolin treatment conditions as for FRA1C. Contrary to my observations at FRA1C, I found that the compaction of this region did not change in response to aphidicolin treatment (Figure 5-3). This reinforced the finding that aphidicolin treatment does not cause genome-wide chromatin compaction and the effect observed at FRA1C is specific either to that region or to an active fragile sites.

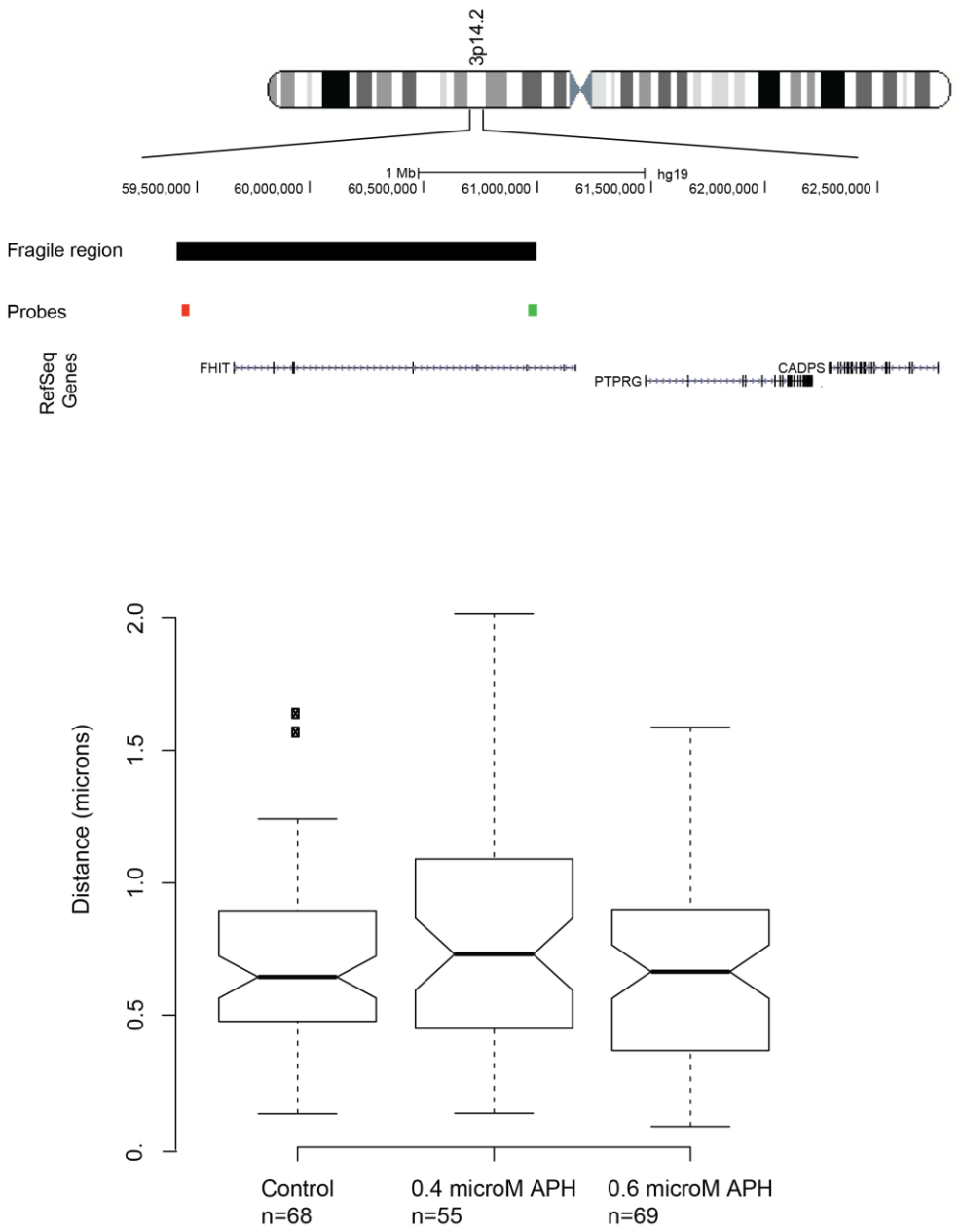


Figure 5-4 Chromatin compaction at the inactive FRA3B locus in RPE1 cells upon aphidicolin treatment. RPE1 cells were grown under control conditions or in the presence of two different concentrations of aphidicolin. Cells were then hybridised to a pair of fosmid probes located 1 Mb apart, which surrounded the FRA3B locus. Probe positions are shown in red and green in the top diagram; the fragile region is shown as a black bar. The bottom graph shows a boxplot of the distance between the two fosmids across the different samples. Numbers of nuclei included in the analysis is shown for each category. Using a Wilcoxon test there are no significant differences between samples

5.2 Interphase chromatin compaction at CFS regions in synchronised cells

Since the investigation of chromatin compaction in asynchronous RPE1 cells indicated that replication stress induced compaction specific to the FRA1C locus, I wanted to examine the cell-cycle dynamics and dependencies of this compaction. Specifically, I wanted to determine if the changes in compaction precede mitotic fragility at the locus and how they relate to replication timing changes in the presence of aphidicolin. To achieve this, I employed an experimental strategy based on arresting the cells at the G1/S boundary, followed by a release, allowing the cell cycle to continue into S phase, G2 and mitosis. At the time of release, cells were either released in drug-free culture media or in the presence of a low dose of aphidicolin, which induces replication stress and CFS fragility. I harvested cells at different times throughout the transition from the G1/S boundary into G2 and assayed chromatin state through FISH at each time point. I performed these experiments in RPE1 cells, monitoring the compaction of FRA1C, as well as in HCT116 cells, in which I focused on the highly fragile FRA3B site. Although analysis of multiple CFS locations would be beneficial, the time-consuming and labour intensive nature of these experiments allowed only two sites to be analysed.

5.2.1 Cell synchronisation

To synchronise cells, I followed a protocol based on arresting cells at the G1/S boundary by adding a high dose of aphidicolin which completely blocks replication (Pedrali-Noy et al. 1980). Aphidicolin is preferred to other agents used for synchronisation at the G1/S boundary, such as hydroxyurea and thymidine, due to the fact that it does not affect the levels of dNTPs and DNA polymerases inside cells, allowing progression into S phase immediately upon release.

Prior to performing the FISH experiments, I wanted to verify that synchronisation was efficient and determine if a low dose of aphidicolin upon release would influence the rate of progression into S phase. I performed the synchronisation

protocol in RPE1 and HCT116 cells and assessed the cell cycle distribution at different time points following release, both when cells were released in non-supplemented media and with release in a low dose of aphidicolin, used to induce replication stress and trigger CFS expression. Cells were pulsed with EdU for 30 minutes prior to harvesting and characterisation of the cell cycle state at each time point was performed by flow cytometry.

In the HCT116 cell line, I found that aphidicolin arrest and release resulted in a synchronised G1/S population which then transitioned through S phase and into G2 within 10 hours (Figure 5-5). Within the first 2 to 6 hours, cells within different S-phase stages were enriched, with early S most prominent at two hours, mid-S at 4 hours and late-S most prominent at 6 hours post release from the aphidicolin block. I found that addition of a small dose of aphidicolin upon release from the G1/S block caused a moderate delay in S-phase progression; surprisingly, the delay appeared more obvious at the later stages of replication, with aphidicolin-treated cells completing replication nearly two hours after non-treated cells (Figure 5-5).

Similar dynamics were observed in the RPE1 cell line (Figure 5-6). Cells appeared to complete replication up to 10 hours following release, with early, mid and late-S phase enrichment at 2, hours, 4 hours and 6/8 hours, respectively. Aphidicolin addition post release also caused a moderate delay in replication progression in the RPE1 cell line.

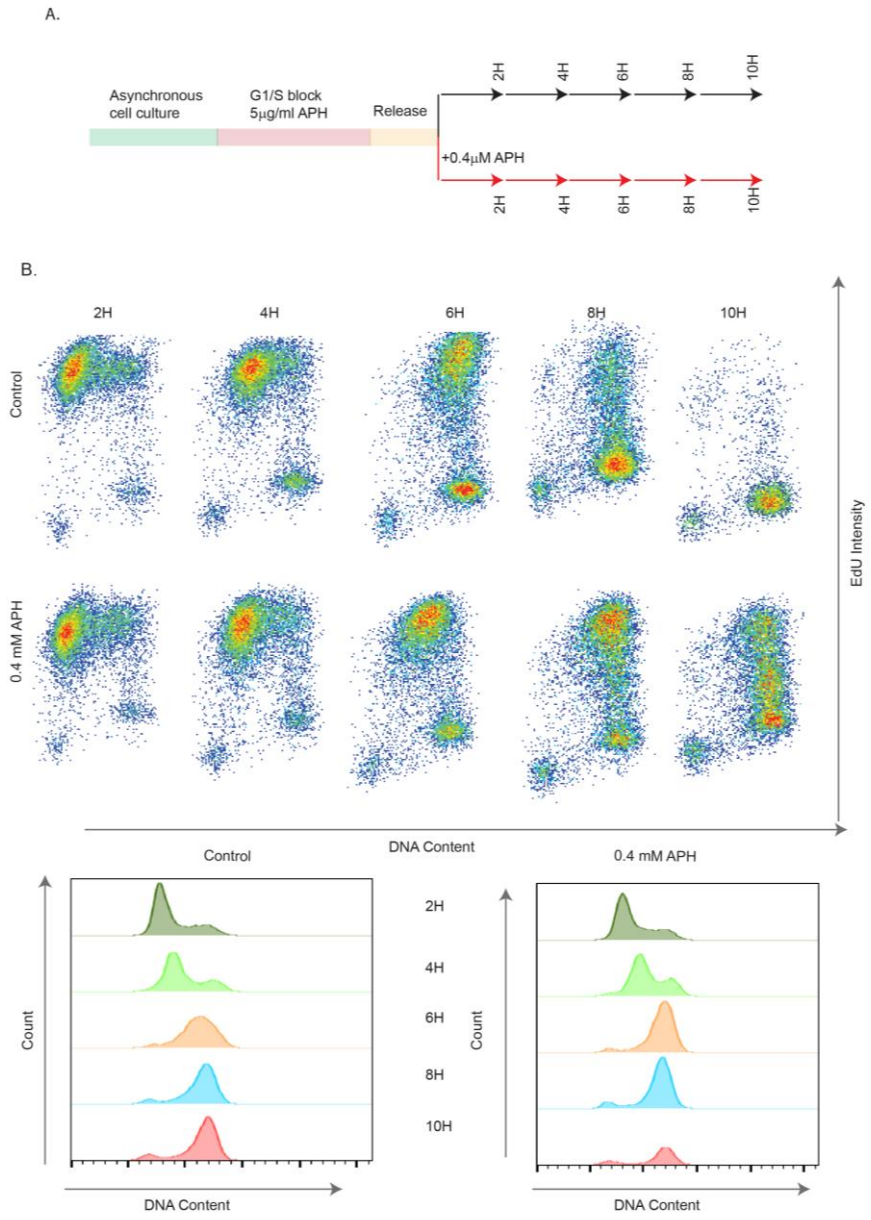


Figure 5-5 Cell synchronisation in the HCT116 cell line. A. Synchronisation protocol. Asynchronously growing cultures were treated with a large dose of aphidicolin for 24 hours, leading to synchronisation at the G1/S boundary. Cells were then released into S phase, either in unperturbed conditions or in the presence of a low dose of aphidicolin, which triggers replication stress and CFS fragility. Progression through S phase and into G2 was tracked by harvesting cells at 2 hour intervals up to 10 hours following release. B. Flow-cytometry based assessment of cell cycle distribution of samples at different time points following release. Samples were pulsed with EdU 30 minutes prior to harvesting to identify replicating cells. Top graphs show EdU intensity versus DNA content, enabling identification of replicating cells. Bottom graphs show PI histograms of the cell populations at different time points.

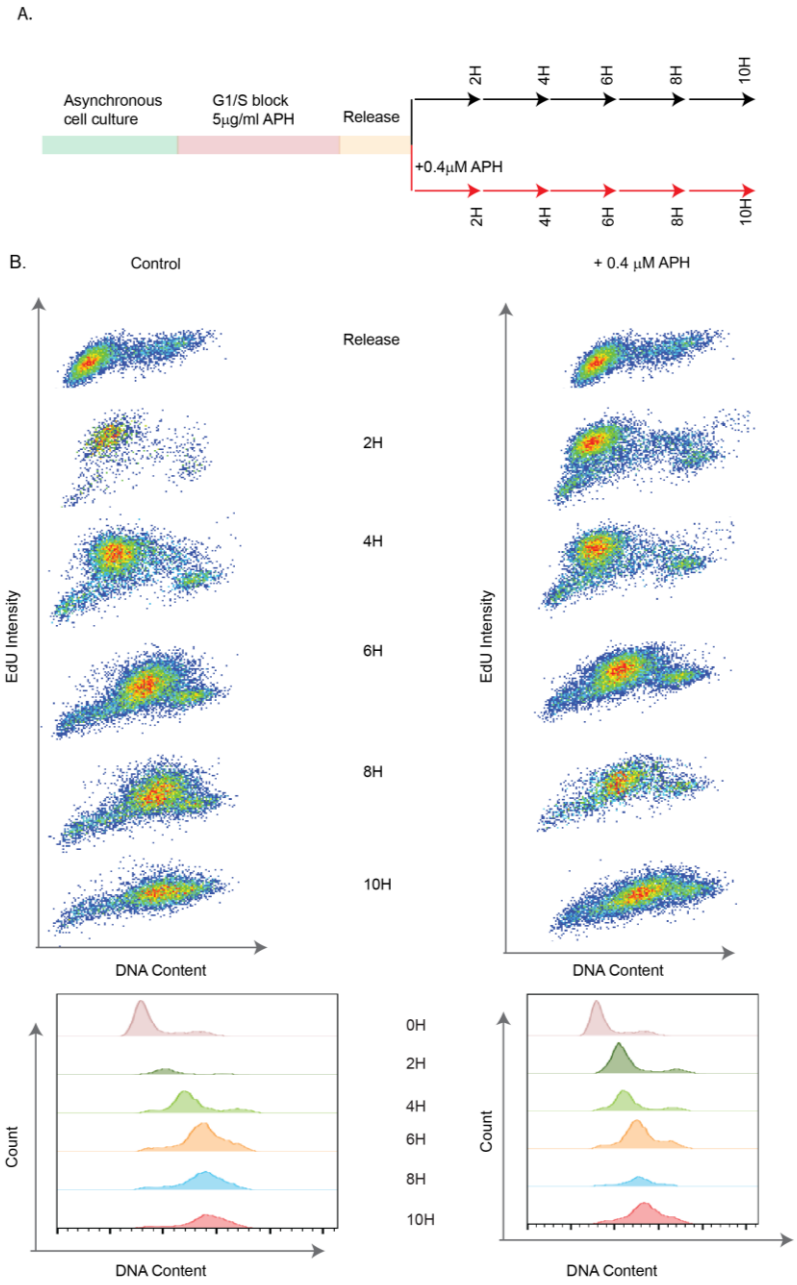


Figure 5-6 Cell synchronisation in the RPE1 cell line. A. Synchronisation protocol. Like in Figure 5-5, asynchronously growing cultures were treated with a large dose of aphidicolin for 24 hours, leading to synchronisation at the G1/S boundary. Following arrest, cells were released into S phase, either in unperturbed conditions or in the presence of low dose aphidicolin. Progression through the cell cycle was tracked by harvesting cells at 2 hour intervals up to 10 hours following release. B. Assessment of cell cycle distribution of samples at different time points following release. Samples were pulsed with EdU 30 minutes prior to harvesting to identify replicating cells. Top graphs show EdU intensity versus DNA content, allowing identification of replicating cells. Bottom graphs show PI histograms of the cell populations at different time points

5.2.2 Chromatin changes at FRA1C throughout the cell cycle

With the cell synchronisation protocol established, I characterised the changes in the large-scale chromatin state of FRA1C in RPE1 cells throughout the transition from G1/S to G2, under unperturbed conditions and in the presence of aphidicolin-induced replication stress.

I prepared cell populations at defined time points following release from the G1/S block and hybridised them to the fosmid pairs within the FRA1C region. I then compared the distances between fosmid pairs across the different time points and in the presence or absence of aphidicolin. At least 60 images were analysed at each time point and each condition. I found that chromatin undergoes significant changes following release from a G1/S block into unperturbed conditions: significant compaction occurs during the transition between release and the two hour time point, coincident with the early replication stage (figure 5-7). The region then significantly decompacts by the four-hour time point, which marks the early to mid-replication stage. Following this decompaction, the chromatin state of the region remains unchanged until the 10-hour time point. Comparison with Click-seq data within the region reveals that a transition between an early/mid and a late replicating region is contained between the fosmid pair; the early/mid replicating region is likely to be replicated by the four hour time point, when the decompaction is observed. Unfortunately, the timing resolution of Click-seq is not sufficient to determine if decompaction precedes replication of the locus or is only established once the site has been replicated.

Chromatin dynamics are subtly different in the presence of aphidicolin. Similar to control cells, the locus significantly compacts between release and the two hour time point. Curiously, unlike the control sample, there was no significant decompaction between the two hour and four hour time point in the presence of aphidicolin, leading to a significant difference in chromatin state between the control sample and the aphidicolin treated sample at the four hour time point. In the presence of aphidicolin, the early/mid portion of the region replicated earlier,

indicating that a change in replication timing was matched by a change in chromatin dynamics at this site. Unlike the asynchronous cell cultures, where it was not possible to determine if the change in chromatin compaction preceded mitotic lesion formation, this experiment demonstrates the change precedes breakage, raising the possibility that chromatin state contributes to CFS instability.

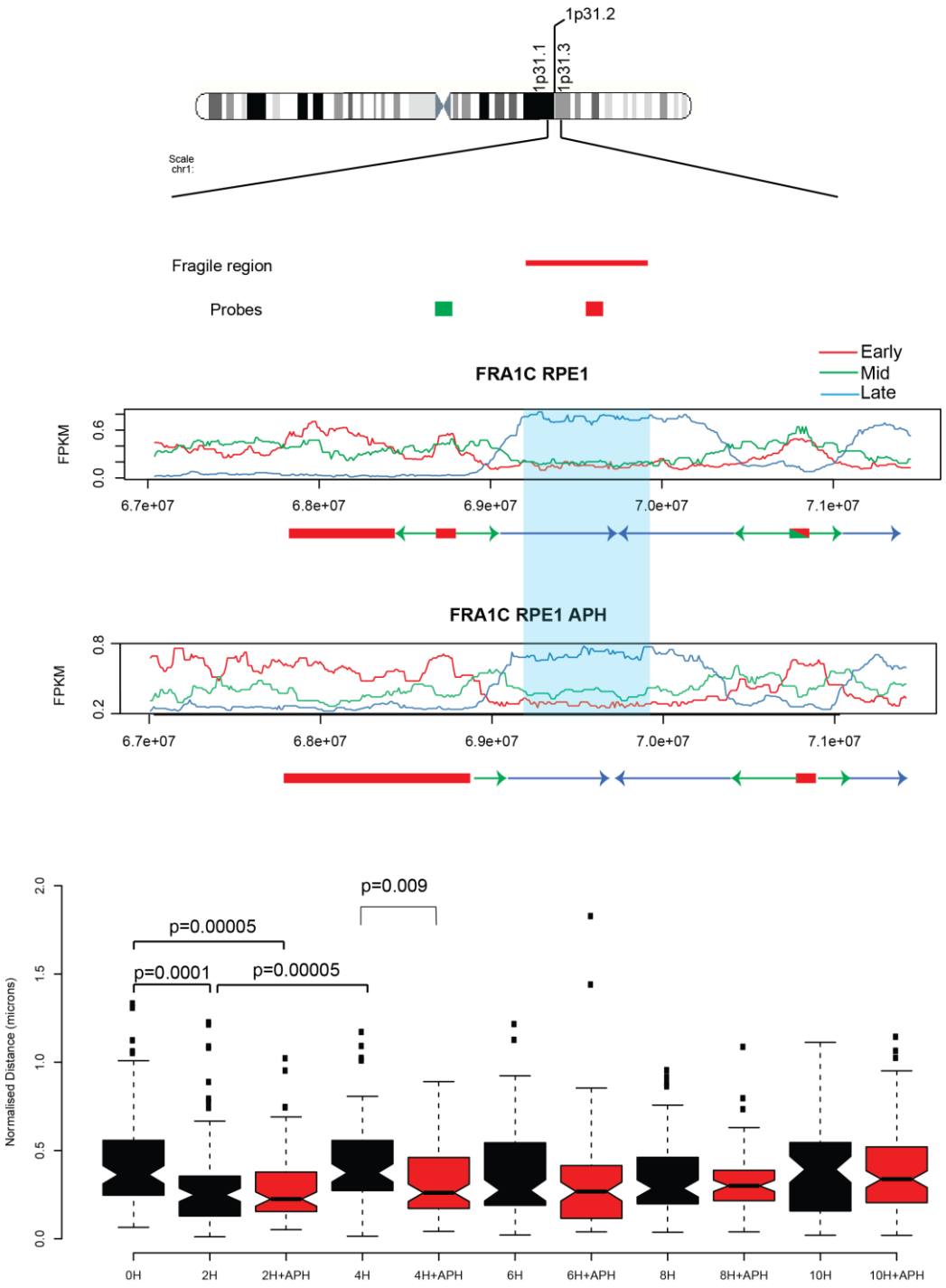


Figure 5-7 Chromatin dynamics throughout the cell cycle at the FRA1C site in RPE1 cells. Top schematic depicts the position of the two fosmid probes relative to the fragile location. Middle panel shows the replication timing across the locus in unperturbed cells and following replication stress, and the fragile location is shaded in blue and schematics represent inferred initiation clusters and fork direction. Bottom panel shows a boxplot of fosmid distances across the different time points; control samples are shown in black and aphidicolin-treated samples are shown in red. P-values from a Wilcoxon test are shown.

5.2.3 Chromatin changes at FRA3B throughout the cell cycle

Unlike FRA1C, no significant changes in chromatin state occurred at the FRA3B locus in the HCT116 cell line. The FRA3B region appeared to have a less compact chromatin state compared to FRA1C – although the probes were separated by a similar distance to probes at FRA1C, the physical distance separating the probes appeared to be larger in that region. Although a small trend for decompaction could be seen as cells transitioned into G2, no statistically significant differences were found between the different categories. The probes at this location span a predominantly late replicating region; aphidicolin treatment seemed to trigger an extra origin within that window and induce a slightly later replication timing. It is possible that the core of that region is replicated so late into the cell cycle that any replication-associated chromatin shifts are not obvious before the final 10 hour time point in the experiment. Also unlike the FRA1C region, aphidicolin did not induce any significant changes in compaction in that region; however, replication dynamics at that site differed significantly from FRA1C. Unlike FRA1C, where a shift towards earlier replication timing occurred within the region framed by the fosmid pairs upon aphidicolin treatment, FRA3B shifted towards later replication timing, including the activation of a putative late origin. The FISH compaction data suggests that either these changes were not accompanied by changes of large-scale chromatin structure or, if any changes happened, they occurred beyond the 10 hour time point.

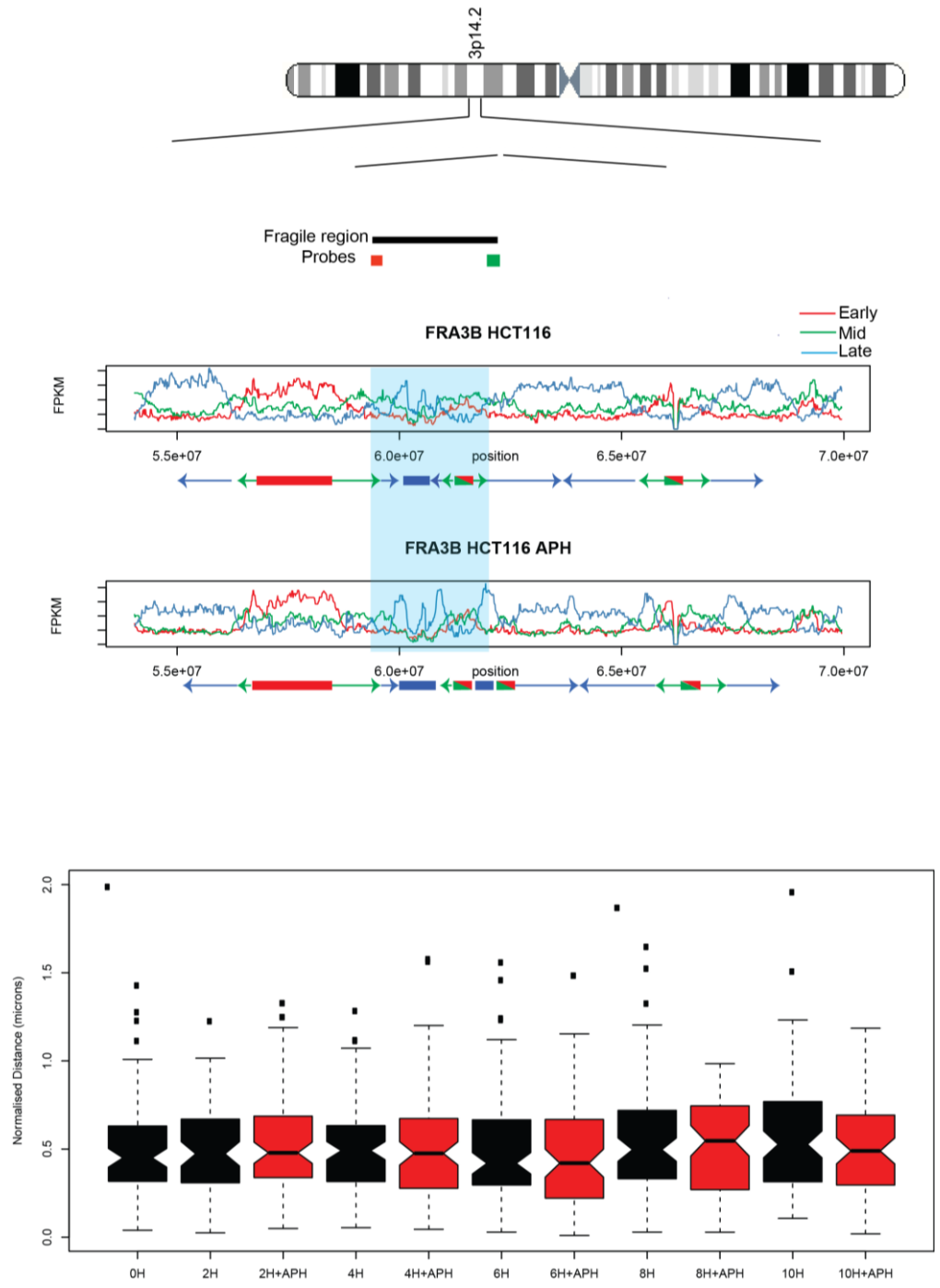


Figure 5-8 Chromatin dynamics throughout the cell cycle at the FRA3B site in HCT116 cells. Top schematic depicts the position of the two fosmid probes relative to the fragile location. Middle panel shows the replication timing across the locus in unperturbed cells and following replication stress, and the fragile location is shaded in blue. Bottom panel shows a boxplot of fosmid distances across the different time points; control samples are shown in black and aphidicolin-treated samples are shown in red.

5.3 Discussion

In this chapter, I have described an investigation of large-scale chromatin structure at two fragile sites: FRA1C and FRA3B, in unperturbed conditions and following replication stress, in asynchronous cultures and in cell populations moving synchronously through S-phase.

Although the fosmid FISH approach has been used extensively before to demonstrate changes in large scale chromatin structure throughout development, X inactivation and upon gene activation, it has never been used to track chromatin dynamics throughout the cell cycle. Cell cycle changes in chromatin structure have been previously analysed in chromosome conformation capture based studies, which showed that the higher order structure of the genome was disrupted in mitosis and re-established early in the following G1 (Naumova et al. 2013; Dileep et al. 2015). No differences were found at the genome-wide level TAD structure throughout S-phase- compartments remained relatively stable and matched the boundaries defined in G1. However, at the chromosome 1p31.2 region (FRA1C), the use of fosmid FISH at defined time points throughout S phase showed that significant changes in large scale structure are under way in this part of the cell cycle. This result suggests that although labour-intensive, the application of fosmid FISH at defined cell cycle stages can yield insights about chromatin dynamics during S-phase. Another finding from the fosmid FISH approach described in this chapter is that replication stress, when induced by aphidicolin, does not cause a genome-wide change in large-scale chromatin structure. This is in contrast to drugs that interfere with transcription, such as alpha-amanitin or compounds that cause disruption in the regulation of histone marks, such as TSA (Naughton et al. 2013; Tóth et al. 2004)

Finally, the use of fosmid FISH allowed me to explore the possibility that chromatin structure contributes to fragility at active CFSs. To investigate that, I characterised chromatin changes in the presence of aphidicolin at two fragile locations in different cell lines: FRA1C in RPE1 and FRA3B in HCT116. Surprisingly, the two locations yielded contrasting observations: while a structural change preceding

mitosis could be observed at FRA1C, no changes were induced by replication stress at FRA3B. Apart from chromatin dynamics, FRA3B and FRA1C differ in many other ways: while FRA3B spans two genes, both of which are expressed in the HCT116 cell line, no expressed transcripts are present at FRA1C. The replication profiles of the two sites also differ: although both sites are replicated late, FRA1C is replicated by forks travelling across from early/mid firing origins, while FRA3B appears to span late-initiating origins. These differences in characteristics and chromatin behaviour may raise the possibility that locus-specific effects are the main contributors to fragility at CFS, rather than shared features. Despite the observation that aphidicolin-induced compaction is an effect specific to FRA1C, it is easy to envisage how it may interfere with the stability of the site by creating a chromatin environment which is less conducive to replication, repair and mitotic compaction at the locus.

6 Chapter 6: Discussion

Common fragile sites are a phenomenon that historically have been observed and defined through cytogenetic means. However, significant efforts to transition from a cytogenetic to a molecular definition of CFS have been made: studies from the Hickson lab suggested that the DNA repair protein FANCD2 can be used as a marker of CFS in G2 and mitosis (Chan et al. 2009) and Le Tallec et al (2013) proposed that genes over 300 kb mark the pool of all potential CFS loci. Another suggested molecular mark limits putative CFSs to late-replicating regions devoid of replication initiation zones and the replication origin protein ORC2 (Miotto et al. 2016; Letessier et al. 2011), while an analysis of cancer mutation catalogues showed that a large number of recurrent cancer deletion clusters are associated with CFSs (Bignell et al. 2010). However, identifying a molecular definition of CFS regions has been impeded by their cell type specific nature and the fact that their fragility appears to be dependent on multiple events across multiple cell cycle stages. In addition, most of the studies proposing a molecular definition of CFS regions have focused on a small number of sites, frequently within a single cell type. Here, I have presented the analysis of multiple CFS features across two cell types with distinct CFS repertoires. While I was unable to define a shared replication timing pattern associated with fragility, or a robust correlation with transcription, I found that CFS regions are characterised by disruptions in chromatin folding on metaphase chromosomes. My data raises the possibility that instability at different sites is driven by different mechanisms, converging into a shared phenotype of mis-folding prior to mitosis and the subsequent formation of chromosome lesions. A summary of all CFS regions investigated in this study and the experiments performed on them is presented in Table 6-1.

CFS Name	Cell type specificity (% of all breaks observed in cell type)		Genomic Location	Analysis performed
FRA1C	RPE1 cells: 18.6%	HCT116 cells: 11.91%	1p31.2	• **Fine mapping with BAC probes** Fragility found to span a 0.6 Mb region around chr1: 69176951- 69781569 • **Mapping of atypical probe signals across the locus** Atypical probe signal distribution found to overlap with fragile region (3.3.3) Replication stress found to increase frequency of atypical signals (3.3.4) • **Replication timing analysis** Site found to be replicated late in S phase, by forks travelling from origins located 0.5 Mb to 1 Mb away; no delay in replication in the presence of aphidicolin (4.5.1.1). • **Analysis of interphase chromatin structure** Chromatin compaction of the locus observed upon aphidicolin treatment in asynchronous cells (5.1.3). Chromatin structure analysis in synchronised cells showed that the change appears during early to mid S phase and precedes breakage at the site (5.2.2).
Novel	RPE1 cells: 11.9	HCT116 cells: Not	4q32.2	• **Fine-mapping with BAC probes** Fragility found to span a 1Mb region overlapping with the MARCH1 gene at 4q32.2-4q32.3 boundary

		fragile		• **Analysis of transcriptional levels within the locus** Increased transcription across the locus in RPE1 cells (3.4.3) • **Replication timing analysis** Site replicated in mid to late S phase by forks travelling from origins located 0.5 Mb to 1 Mb away. Replication across the site is delayed in the presence of aphidicolin (4.5.1.1).
FRA3B	RPE1 cells: Not fragile	HCT116 cells: 20%	3p14.2	• **Fine-mapping with fosmid probes** Fragility localised to a 1 Mb region overlapping with the FHIT gene at 3p14.2 • **Analysis of transcriptional levels within the locus** Increased transcription across the locus in HCT116 cells (3.4.3) • **Modifying transcription levels at the site using CRISPR Cas9 (3.5)** An small increase in the frequency of lesions upon reduction of FHIT transcription A larger increase in the frequency of lesions upon an increase in FHIT transcription • **Replication timing analysis** Site is late-replicating and overlaps late-firing origins. Replication at the site is delayed in the presence of aphidicolin and extra origin is fired (4.5.1.2). • **Analysis of interphase chromatin structure** No changes in chromatin structure at the locus upon aphidicolin treatment in asynchronous cells

				(5.1.4) or synchronised cell populations (5.2.3).
FRA4F	RPE1 cells: Not fragile	HCT116 cells: 13.7%	4q22.2	• **Fine-mapping with BAC probes** Fragility localised to a 5 Mb region between chr4: 89213170-95121208 • **Mapping of atypical FISH probe signals across the locus** Atypical probe signal distribution found to overlap with fragile region (3.3.2) Replication stress found to increase frequency of atypical signals (3.3.4) Calyculin experiment to analyse dynamics of mitotic folding at the locus (3.3.5) • **Analysis of transcriptional levels within the locus** Increased transcription across the locus in RPE1 cells (3.4.3) • **Replication timing analysis** Site spans and early to mid-replicating domain and a late replicating domain (4.5.1.2).
FRA2F	RPE1 8.5%	HCT116 8.7%	2q22.2	• **Fine-mapping with BAC probes** Fragility localised to a 2 Mb region between chr2: 142200000-144100000 • **Analysis of transcriptional levels within the locus** No difference in transcription across the locus between the two cell lines (3.4.3) • **Replication timing analysis** Site is predominantly late-replicating. Aphidicolin treatment results in earlier replication timing (4.5.1.2).

FRA3O	RPE1 cells: 16.9 %	HCT116 cells: 5.5%		• **Analysis of transcriptional levels within the locus** Increased transcription across the locus in RPE1 cells (3.4.3)
Novel	RPE1 cells: 10.2%	HCT116 cells: Not fragile	7q21.22	• **Analysis of transcriptional levels within the locus** Increased transcription across the locus in RPE1 cells (3.4.3)
FRA2I	RPE1 cells: Not fragile	HCT116 cells: 15%		• **Analysis of transcriptional levels within the locus** Increased transcription across the locus in RPE1 cells (3.4.3)
FRA2T	RPE1 cells: Not fragile	HCT116 cells: 15%		• **Analysis of transcriptional levels within the locus** Increased transcription across the locus in RPE1 cells (3.4.3)

Table 6-1 Sumamry of CFS studied

6.1 Replication and CFS

The fragility of CFS regions has long been thought to be rooted in the process of replication. A simple replication-based model for fragility is built on the idea that CFS span regions devoid of replication initiation zones and are instead replicated by long-travelling forks originating outside the fragile regions. In this model, replication stress causes a slow-down in fork speed, which interferes with the complete

replication of the site. This mechanism was proposed after the observation that the FRA3B region does not span an initiation cluster in lymphoblastoid cell lines (Letessier et al. 2011) and is supported by a recent study mapping the genome-wide occupancy of the origin recognition complex component ORC2, which found that initiation-poor zones overlap with CFS (Miotto et al. 2016). Across the sites investigated in this study, some CFSs, such as FRA1C and the novel chr4q32.2-32.3 site, matched that pattern, while others, such as FRA3B and FRA4F, span late-firing origins and contained very small regions that appeared under-replicated. Disappointingly, no striking similarities in replication patterns were found for the CFS regions included in the study and the genomic regions considered did not show extreme replication timing changes in response to aphidicolin treatment. One possible explanation may be that lesion formation at CFS sites is a relatively rare event: even at FRA3B, the most common CFSs in this study, breaks occurred in only 18% of metaphases. As Click-seq measures replication dynamics across the whole cell population, it is possible that replication events leading to fragility are too rare to define through this method. For example, a failure to initiate the late origins at FRA3B and FRA4F in a small subset of cells may not be obvious at the population level. Another possibility is that temporal replication dynamics are not as strong a determinant of CFS behaviour as proposed previously: for example, it has been shown that the FRA3B sequence retained fragility even when integrated in a genomic location with an earlier replication timing (Ragland et al. 2008) and FRA7H was shown to span a transition between an early and a late replicating region (Hellman et al. 2000). However, as the majority of sites identified in this study were late-replicating, the Click-seq data supports late replication timing as a significant, but not complete determinant of instability at CFS.

6.2 CFS in mitosis: structure and function

In Chapter 3.3.1, I demonstrated that FISH probes hybridised to active CFS regions showed atypical signals, consistent with disruptions in the metaphase chromosome structure. Interestingly, similar FISH signals are associated with the fragile telomere phenotype, observed upon APH treatment or upon depletion of TRF1, a component

of the shelterin complex (Sfeir et al. 2009). Both phenotypes are indicative of defective chromosome condensation. At telomeres, the phenotype is thought to result from replication problems such as fork collapses and G-quadruplex structures formed by the GC-rich telomeric repeats; however, CFSs are devoid of repetitive sequences and it is unclear how small-scale events such as fork collapses can lead to fragility and failure of mitotic compaction on such a large genomic scale. At CFS regions, there are two yet unknown questions relating to the phenotype: what are the mechanisms leading to mis-folding and how does it relate to the lesions historically observed at the sites. A recent study by the Hickson lab found that DNA synthesis by POLD3 can be observed at CFS in mitosis following aphidicolin treatment and speculated that under-replicated regions are "exposed" by the forces of mitotic compaction, allowing this synthesis to take place (Minocherhomji et al. 2015). Given that condensin localises to replicated regions of the genome, coupling replication to mitotic compaction, it is highly likely that the misfolding phenotype is a consequence of replication problems and a possible precursor of mitotic repair synthesis. Consistently, premature chromosome condensation experiments described in Chapter 3.3.5 showed that the atypical signals at CFSs are a result of a failure to prepare the chromatin environment for mitosis.

An interesting characteristic of the misfolding phenotype is that it also occurred on cytogenetically normal chromosomes, suggesting that CFS regions may experience problems with mitotic compaction more frequently than indicated by the formation of cytogenetically visible breaks. In the case of HCT116 cells, the misfolding was even present at a low level in the absence of replication stress. This is complementary with another long-standing observation on CFS: aphidicolin treatment and the subsequent lesion formation do not trigger cycle checkpoint activation. Intuitively, breaks, gaps and constrictions at CFS regions appear pathogenic, however the idea that they are just a stage of the processing of these regions has been gaining support. This is substantiated by findings that depletion of the nuclease Mus81 results in a decreased appearance of CFS breaks on mitotic chromosomes, but an increase in DNA damage in the subsequent G1 (Naim et al.

2013). Inhibition of mitotic DNA repair also resulted in increased DNA damage in the following G1 (Minocherhomji et al. 2015). These observations may suggest the presence of a dedicated pathway for resolving mitotic misfolding, operating through Mus81 and mitotic DNA synthesis, with cytogenetic lesions marking areas prone to mis-folding in mitosis and representing a processing intermediate.

6.3 Consequences of replication stress

Replication stress is a phenomenon with a significant physiological relevance in the field of cancer biology. It is well established that replication stress accompanies early tumour development and promotes genomic instability; oncogene activation has been shown to trigger cells into replication with an insufficient nucleotide pool and induce fragile site expression (Bester et al. 2011; Miron et al. 2015). In addition, the origin checkpoint in G1 which ensures a sufficient number of origins are licensed, can be impaired in cancer cells (Shreeram et al. 2002), meaning cells can enter S phase with reduced origin numbers. Consistently, the replication landscape of the tumour-derived HCT116 cell line appeared to have larger, more contiguous domains than the non-tumorigenic RPE1 line, compatible with fewer initiation clusters and the forks travelling longer distances between initiation zones. Reduced origin licensing and increased inter-origin distance in the HCT116 cell line may also explain the cell line's increased sensitivity to aphidicolin, as indicated by the higher rate of CFS breakage and mitotic misfolding.

While changes in fork speed and origin activation rates in condition of replication stress have been thoroughly investigated through DNA fibre FISH and other methods, the Click-seq experiments described in Chapter 4 provide new information on genome-wide changes in replication timing in such conditions. The most surprising outcome of aphidicolin treatment was the induction of bi-directional changes in replication timing, rather than a universal delay as expected from evidence that aphidicolin causes a slow-down in fork speed. Especially surprising was the observation that some origin clusters fired earlier in the presence of aphidicolin than in the control sample. However, previous studies have indicated

that the speed of replication fork can affect origin firing and extra origins can be recruited upon fork slow down, potentially causing some genomic locations to replicate earlier. A study from the Debatisse lab even indicated that effects of replication stress can be traced into the subsequent G1, with additional origins licensed compared to cells not exposed to replication stress (Courbet et al. 2008). As cells in the Click-seq experiments were treated with aphidicolin for 24 hours prior to the EdU pulse and harvest, it is possible that earlier firing of some origins may be the result of replication stress in the previous S-phase and illustrate the consequences of on-going replication stress. It is intriguing to speculate that earlier replication timing of a region may lead to changes in the composition of chromatin assembled at the site post-replication, ultimately relating to some of the epigenetic changes characteristic of tumour cells.

A major drawback of the Click-seq technique is that it cannot define origin locations unambiguously. This is especially problematic for late replicating regions: while early initiation clusters are easy to infer, peaks in the late replicating sample could indicate both termination zones where forks converge with high frequency or a firing of a late cluster. Click-seq data would be maximally informative if combined with information on changes in origin firing under conditions of replication stress. While techniques designed to map origins in mammalian cells were previously deemed somehow unreliable, a newly developed methodology, OK-seq, appears more robust and informative (Petryk et al. 2015). OK-seq is relatively easy to implement, particularly for our lab where related techniques are already setup, enabling the mapping of origins in multiple cell lines and under multiple conditions.

Chapter 5 showed that at some sites large-scale structural changes can accompany replication timing changes, illustrating some of the unsuspected consequences of replication stress. Given that the partition of replication timing is related to the 3D organisation of the genome (Dileep et al. 2015) and that replication stress can have knock-on effects on chromatin loop size in the following cell cycle (Courbet et al.

2008), chromosome conformation capture experiments in conditions of replication stress may also provide valuable insights.

6.4 Perspectives

Cytogenetic CFS breaks on metaphase chromosomes have fascinated biologists for a long time. As the understanding of the underlying mechanisms deepens, it becomes clearer that they are a result of complex dependencies between cellular processes such as transcription, replication, and chromatin dynamics. As regions uniquely responsive to replication stress, these sites also represent a valuable model of genomic instability. In this thesis, I have shown a previously unknown tendency for CFS to form regions of misfolding in mitotic chromosomes, which fits with the current understanding of mitotic dynamics and processes of these sites. I have also defined for the first time the cell type specific replication programmes across these regions and characterised how they change in the presence of replication stress. I conclude that rather than sharing a single replication pattern, problems at CFS arise as a result of a variety of replication configurations, converging in their tendency to fail to prepare the chromatin environment for mitosis. My findings contribute towards a more complete view of CFS regions, how they fit within the genomic context and help to inform our view of how the genome is shaped by underlying cellular processes.

7 Chapter 7: References

El Achkar, E. et al., 2005. Premature condensation induces breaks at the interface of early and late replicating chromosome bands bearing common fragile sites. *Proc Natl Acad Sci U S A*, 102, pp.18069–18074.

Adachi, Y., Luke, M. & Laemmli, U.K., 1991. Chromosome assembly in vitro: Topoisomerase II is required for condensation. *Cell*, 64(1), pp.137–148.

Adolph, K.W., Cheng, S.M. & Laemmli, U.K., 1977. Role of nonhistone proteins in metaphase chromosome structure. *Cell*, 12(3), pp.805–816.

Ahmad, K. & Henikoff, S., 2002. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Molecular Cell*, 9, pp.1191–1200.

Alabert, C. et al., 2014. Nascent chromatin capture proteomics determines chromatin dynamics during DNA replication and identifies unknown fork components. *Nature cell biology*, 16, pp.281–93.

Alabert, C. et al., 2015. Two distinct modes for propagation of histone PTMs across the cell cycle. *Genes and Development*, 29(6), pp.585–590.

Allan, J. et al., 1980. The structure of histone H1 and its location in chromatin. *Nature*, 288(5792), pp.675–679.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–10.

Arlt, M.F. et al., 2009. Replication Stress Induces Genome-wide Copy Number Changes in Human Cells that Resemble Polymorphic and Pathogenic Variants. *American Journal of Human Genetics*, 84(3), pp.339–350.

Ashour, M.E., Atteya, R. & El-Khamisy, S.F., 2015. Topoisomerase-mediated chromosomal break repair: an emerging player in many games. *Nature Reviews Cancer*, 15(3), pp.137–151.

Aymard, F. et al., 2014. Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. *Nature structural & molecular biology*, 21(4), pp.366–74.

Baranovskiy, A.G. et al., 2014. Structural basis for inhibition of DNA replication by aphidicolin. *Nucleic Acids Research*, 42(22), pp.14013–14021.

Barber, T.D. et al., 2008. Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9), pp.3443–8.

Baxter, J. & Diffley, J.F.X., 2008. Topoisomerase II Inactivation Prevents the Completion of DNA Replication in Budding Yeast. *Molecular Cell*, 30(6), pp.790–802.

Le Beau, M., 1998. Replication of a common fragile site, FRA3B, occurs late in S phase and is delayed further upon induction: implications for the mechanism of fragile site induction. *Human Molecular Genetics*, 7(4), pp.755–761.

Becker, N. a et al., 2002. Evidence that instability within the FRA3B region extends four megabases. *Oncogene*, 21(57), pp.8713–8722.

Belmont, A.S., Sedat, J.W. & Agard, D.A., 1987. A three-dimensional approach to mitotic chromosome structure: evidence for a complex hierarchical organization. *The Journal of cell biology*, 105(1), pp.77–92.

Benjamini, Y. & Speed, T.P., 2012. RSeQC: Quality Control of RNA-seq experiments. *Bioinformatics (Oxford, England)*, 40(10), p.e72.

Bester, A.C. et al., 2011. Nucleotide deficiency promotes genomic instability in early stages of cancer development. *Cell*, 145, pp.435–446.

Bickmore, W.A., 2013. The spatial organization of the human genome. *Annu Rev Genomics Hum Genet*, 14, pp.67–84.

Bickmore, W.A. & van Steensel, B., 2013. Genome architecture: domain organization of interphase chromosomes. *Cell*, 152, pp.1270–1284.

Bickmore, W.A. & Teague, P., 2002. Influences of chromosome size, gene density and nuclear position on the frequency of constitutional translocations in the human population. *Chromosome Research*, 10(8), pp.707–715.

Bignell, G.R. et al., 2010. Signatures of mutation and selection in the cancer genome. *Nature*, 463(7283), pp.893–898.

Bischof, O. et al., 2001. Regulation and localization of the Bloom syndrome protein in response to DNA damage. *J Cell Biol*, 153(2), pp.367–380.

Blow, J.J., Ge, X.Q. & Jackson, D.A., 2011. How dormant origins promote complete genome replication. *Trends in Biochemical Sciences*, 36(8), pp.405–414.

Boteva, L. & Gilbert, N., 2016. Chromatin, nculear organisation and genome stability in mammals". *Genome Stability, Kovalchuk and Kovalchuk*

Boyle, S. et al., 2011. Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. *Chromosome Research*, 19(7), pp.901–909.

Boyle, S. et al., 2001. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, 10(3), pp.211–219.

Britton, S. et al., 2014. DNA damage triggers SAF-A and RNA biogenesis factors exclusion from chromatin coupled to R-loops removal. *Nucleic acids research*, pp.1–16.

Brownlee, P.M. et al., 2014. BAF180 Promotes Cohesion and Prevents Genome Instability and Aneuploidy. *Cell Reports*, 6(6), pp.973–981.

Burman, B. et al., 2015. Histone modifications predispose genome regions to breakage and translocation. *Genes & Development*, 29(13), pp.1393–1402.

Burrell, R.A. et al., 2013. Replication stress links structural and numerical cancer chromosomal instability. *Nature*, 494, pp.492–496.

Carpenter, A.J., 2004. Construction, Characterization, and Complementation of a Conditional-Lethal DNA Topoisomerase II Mutant Human Cell Line. *Molecular Biology of the Cell*, 15(12), pp.5700–5711.

Casper, A.M. et al., 2004. Chromosomal instability at common fragile sites in Seckel syndrome. *Am J Hum Genet*, 75, pp.654–660.

Castedo, M. et al., 2004. Cell death by mitotic catastrophe: a molecular definition. *Oncogene*, 23(16), pp.2825–2837.

Celeste, A. et al., 2003. H2AX haploinsufficiency modifies genomic stability and tumor susceptibility. *Cell*, 114(3), pp.371–383.

Chambeyron, S. & Bickmore, W.A., 2004. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes and Development*, 18(10), pp.1119–1130.

Chan, K.L. et al., 2009. Replication stress induces sister-chromatid bridging at fragile site loci in mitosis. *Nat Cell Biol*, 11, pp.753–760.

Chen, B. et al., 2013. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell*, 155(7), pp.1479–1491.

Chowdhury, D. et al., 2005. gamma-H2AX dephosphorylation by protein phosphatase 2A facilitates DNA double-strand break repair. *Molecular cell*, 20(5), pp.801–809.

Clemson, C.M. et al., 1996. XIST RNA paints the inactive X chromosome at interphase: Evidence for a novel RNA involved in nuclear/chromosome structure. *Journal of Cell Biology*, 132(3), pp.259–275.

Constantinou, A. et al., 2000. Werner's syndrome protein (WRN) migrates Holliday junctions and co-localizes with RPA upon replication arrest. *EMBO reports*, 1(1), pp.80–4.

Corona, D.F.V. & Tamkun, J.W., 2004. Multiple roles for ISWI in transcription, chromosome organization and DNA replication. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1677(1-3), pp.113–119.

Courbet, S. et al., 2008. Replication fork movement sets chromatin loop size and origin choice in mammalian cells. *Nature*, 455(iii), pp.557–560.

Courilleau, C. et al., 2012. The chromatin remodeler p400 atpase facilitates RAD51-mediated repair of DNA double-strand breaks. *Journal of Cell Biology*, 199(7), pp.1067–1081.

Craig, J.M. et al., 1997. Scaffold attachments within the human genome. *Journal of cell science*, 110 (Pt 2, pp.2673–2682.

Craig-Holmes, A.P. et al., 1987. Variation in the expression of aphidicolin-induced fragile sites in human lymphocyte cultures. *Human Genetics*, 76(2), pp.134–137.

Croft, J.A. et al., 1999. Differences in the localization and morphology of chromosomes in the human nucleus. *Journal of Cell Biology*, 145(6), pp.1119–1131.

Curanovic, D. et al., 2013. Global profiling of stimulus-induced polyadenylation in cells using a poly(A) trap. *Nat Chem Biol*, 9, pp.671–673.

Cusick, M.E. et al., 1983. Structure of chromatin at deoxyribonucleic acid replication forks: nuclease hypersensitivity results from both prenucleosomal deoxyribonucleic acid and an immature chromatin structure. *Biochemistry*, 22(16), pp.3873–84.

Debatisse, M., El Achkar, E. & Dutrillaux, B., 2006. Common fragile sites nested at the interfaces of early and late-replicating chromosome bands: Cis acting components of the G2/M checkpoint? *Cell Cycle*, 5(6), pp.578–581.

Dellaire, G., Kepkay, R. & Bazett-Jones, D.P., 2009. High resolution imaging of changes in the structure and spatial organization of chromatin, gamma-H2A.X and the MRN complex within etoposide-induced DNA repair foci. *Cell cycle*, 8, pp.3750–3769.

Dileep, V. et al., 2015. Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program. *Genome Research*, 25(8), pp.1104–1113.

Dileep, V., Didier, R. & Gilbert, D.M., 2012. Genome-wide analysis of replication timing in mammalian cells: troubleshooting problems encountered when comparing different cell types. *Methods (San Diego, Calif.)*, 57(2), pp.165–9.

Dimitrova, D.S. & Berezney, R., 2002. The spatio-temporal organization of DNA replication sites is identical in primary, immortalized and transformed mammalian cells. *J Cell Sci*, 115, pp.4037–4051.

Dimitrova, D.S. & Gilbert, D.M., 1999. The spatial position and replication timing of chromosomal domains are both established in early G1 phase. *Mol Cell*, 4, pp.983–993.

Dovey, O.M. et al., 2013. Histone deacetylase 1 and 2 are essential for normal T-cell development and genomic stability in mice. *Blood*, 121, pp.1335–1344.

Dulev, S., Aragon, L. & Strunnikov, A., 2008. Unreplicated DNA in mitosis precludes condensin binding and chromosome condensation in S. cerevisiae. *Front Biosci*, 13, pp.5838–5846.

Durkin, S.G. & Glover, T.W., 2007. Chromosome fragile sites. *Annu Rev Genet*, 41, pp.169–192.

Dykhuizen, E.C. et al., 2013. BAF complexes facilitate decatenation of DNA by topoisomerase IIα. *Nature*, 497, pp.624–627.

Engh, G., Sachs, R. & Trask, B.J., 1992. Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science (New York, N.Y.)*, 257(5075), pp.1410–1412.

Fernandez-Capetillo, O., Allis, C.D. & Nussenzweig, A., 2004. Phosphorylation of histone H2B at DNA double-strand breaks. *The Journal of experimental medicine*, 199, pp.1671–1677.

Francia, S. et al., 2012. Site-specific DICER and DROSHA RNA products control the DNA-damage response. *Nature*, 488(7410), pp.231–235.

Foti, R. et al., 2016. Nuclear Architecture Organized by Rif1 Underpins the Replication-Timing Program. *Mol Cell* 61(2), pp260-273

Fungtammasan, A. et al., 2012. A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? *Genome Res*, 22, pp.993–1005.

Gaillard, H., García-Muse, T. & Aguilera, A., 2015. Replication stress and cancer. *Nature Reviews Cancer*, 15(5), pp.276–289.

Gebhart, E. et al., 1988. Spontaneous and induced chromosomal instability in Werner syndrome. *Hum Genet*, 80(2), pp.135–139.

Gilbert, N. & Allan, J., 2001. Distinctive higher-order chromatin structure at mammalian centromeres. *Proceedings of the National Academy of Sciences of the United States of America*, 98, pp.11949–11954.

Gilchrist, S. et al., 2004. Nuclear organization of centromeric domains is not perturbed by inhibition of histone deacetylases. *Chromosome Research*, 12(5), pp.505–516.

Gimenez-Abian, J.F. et al., 1995. A postprophase topoisomerase II-dependent chromatid core separation step in the formation of metaphase chromosomes. *Journal of Cell Biology*, 131(1), pp.7–17.

Giraud, F. et al., 1976. Constitutional chromosomal breakage. *Hum Genet*, 34, pp.125–136.

Gong, F. & Miller, K.M., 2013. Mammalian DNA repair: HATs and HDACs make their mark through histone acetylation. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 750, pp.23–30.

Goodarzi, A.A. et al., 2008. ATM Signaling Facilitates Repair of DNA Double-Strand Breaks Associated with Heterochromatin. *Molecular Cell*, 31, pp.167–177.

Goodarzi, A.A., Kurka, T. & Jeggo, P.A., 2011. KAP-1 phosphorylation regulates CHD3 nucleosome remodeling during the DNA double-strand break response. *Nature structural & molecular biology*, 18(7), pp.831–839.

Gospodinov, A. et al., 2011. Mammalian Ino80 Mediates Double-Strand Break Repair through Its Role in DNA End Strand Resection. *Molecular and Cellular Biology*, 31(23), pp.4735–4745.

Green, C.M. & Almouzni, G., 2002. When repair meets chromatin. First in series on chromatin dynamics. *EMBO Reports*, 3(1), pp.28–33.

Green, L.C. et al., 2012. Contrasting roles of condensin I and condensin II in mitotic chromosome formation. *J Cell Sci*, 125, pp.1591–1604.

Grigoryev, S. a & Woodcock, C.L., 2012. Chromatin organization - the 30 nm fiber. *Experimental cell research*, 318(12), pp.1448–55.

Guelen, L. et al., 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197), pp.948–51.

Haering, C.H. et al., 2002. Molecular architecture of SMC proteins and the yeast cohesin complex. *Molecular Cell*, 9(4), pp.773–788.

Hall, L.L. et al., 2014. Stable C0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. *Cell*, 156(5), pp.907–919.

Hamilton, C., Hayward, R.L. & Gilbert, N., 2011. Global chromatin fibre compaction in response to DNA damage. *Biochemical and Biophysical Research Communications*, 414, pp.820–825.

Hansen, R.S. et al., 2010. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1), pp.139–44.

Harrigan, J.A. et al., 2011. Replication stress induces 53BP1-containing OPT domains in G1 cells. *J Cell Biol*, 193, pp.97–108.

Harshman, S.W. et al., 2013. H1 histones: Current perspectives and challenges. *Nucleic Acids Research*, 41(21), pp.9593–9609.

Hellman, A. et al., 2000. Replication delay along FRA7H, a common fragile site on human chromosome 7, leads to chromosomal instability. *Molecular and cellular biology*, 20(12), pp.4420–7.

Helmrich, A. et al., 2006. Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes. *Genome Res*, 16, pp.1222–1230.

Helmrich, A., Ballarino, M. & Tora, L., 2011. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol Cell*, 44, pp.966–977.

Henikoff, S., Furuyama, T. & Ahmad, K., 2004. Histone variants, nucleosome assembly and epigenetic inheritance. *Trends in Genetics*, 20(7), pp.320–326.

Heo, K. et al., 2008. FACT-Mediated Exchange of Histone Variant H2AX Regulated by Phosphorylation of H2AX and ADP-Ribosylation of Spt16. *Molecular Cell*, 30(1), pp.86–97.

Hiratani, I. et al., 2009. Replication timing and transcriptional control: beyond cause and effect--part II. *Curr Opin Genet Dev*, 19, pp.142–149.

Hu, Y. et al., 2009. Recql5 plays an important role in DNA replication and cell survival after camptothecin treatment. *Molecular biology of the cell*, 20(1), pp.114–123.

Huang, H. et al., 1998. FRA7G extends over a broad region: coincidence of human endogenous retroviral sequences (HERV-H) and small polydispersed circular DNAs (spcDNA) and fragile sites. *Oncogene*, 16(18), pp.2311–2319.

Hur, S.K. et al., 2010. Roles of human INO80 chromatin remodeling enzyme in DNA replication and chromosome segregation suppress genome instability. *Cellular and Molecular Life Sciences*, 67(13), pp.2283–2296.

Huyen, Y. et al., 2004. Methylated lysine 79 of histone H3 targets 53BP1 to DNA double-strand breaks. *Nature*, 432, pp.406–411.

Jasencakova, Z. et al., 2010. Replication Stress Interferes with Histone Recycling and Predeposition Marking of New Histones. *Molecular Cell*, 37(5), pp.736–743.

Jiang, Y. et al., 2009. Common fragile sites are characterized by histone hypoacetylation. *Hum Mol Genet*, 18, pp.4501–4512.

Kakarougkas, A. et al., 2014. Requirement for PBAF in transcriptional repression and repair at DNA breaks in actively transcribed regions of chromatin. *Mol Cell*, 55, pp.723–732.

Kanda, R., Eguchi-Kasai, K. & Hayata, I., 1999. Phosphatase inhibitors and premature chromosome condensation in human peripheral lymphocytes at different cell-cycle phases. *Somatic cell and molecular genetics*, 25(1), pp.1–8.

Kent, W.J. et al., 2002. The Human Genome Browser at UCSC. *Genome Research*, 12(6), pp.996–1006.

Kim, J.S. et al., 2002. Specific recruitment of human cohesin to laser-induced DNA damage. *Journal of Biological Chemistry*, 277(47), pp.45149–45153.

King, I.F. et al., 2013. Topoisomerases facilitate transcription of long genes linked to autism. *Nature*, 501(7465), pp.58–62.

Kon, A. et al., 2013. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nature Genetics*, 45(10), pp.1232–1237.

Krokan, H., Wist, E. & Krokan, R.H., 1981. Aphidicolin inhibits DNA synthesis by DNA polymerase alpha and isolated nuclei by a similar mechanism. *Nucleic acids research*, 9(18), pp.4709–19.

Kruhlak, M.J. et al., 2006. Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks. *The Journal of cell biology*, 172(6), pp.823–834.

Kusch, T. et al., 2004. Acetylation by Tip60 is required for selective histone variant exchange at DNA lesions. *Science (New York, N.Y.)*, 306, pp.2084–2087.

Lan, L. et al., 2010. The ACF1 Complex Is Required for DNA Double-Strand Break Repair in Human Cells. *Molecular Cell*, 40(6), pp.976–987.

Langmead, B., 2010. Aligning short sequencing reads with Bowtie. *Current Protocols in Bioinformatics*, (SUPP.32).

Larsen, D.H. et al., 2010. The chromatin-remodeling factor CHD4 coordinates signaling and repair after DNA damage. *Journal of Cell Biology*, 190(5), pp.731–740.

Lee, J.-H. et al., 2015. AKT phosphorylates H3-threonine 45 to facilitate termination of gene transcription in response to DNA damage. *Nucleic Acids Research*, 43(9), pp.4505–4516.

Lee, J.-S., 2007. Activation of ATM-dependent DNA damage signal pathway by a histone deacetylase inhibitor, trichostatin A. *Cancer research and treatment : official journal of Korean Cancer Association*, 39, pp.125–130.

Letessier, A. et al., 2011. Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. *Nature*, 470, pp.120–123.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.

Liang, Z. et al., 2015. Chromosomes Progress to Metaphase in Multiple Discrete Steps via Global Compaction/Expansion Cycles. *Cell*, 161(5), pp.1124–1137.

Liu, J. et al., 2009. Transcriptional dysregulation in NIPBL and cohesin mutant human cells. *PLoS biology*, 7(5), p.e1000119.

Lubelsky, Y. et al., 2014. DNA replication and transcription programs respond to the same chromatin cues. *Genome Research*, 24(7), pp.1102–1114.

Ma, N. et al., 2011. The nuclear scaffold protein SAF-A is required for kinetochore-microtubule attachment and contributes to the targeting of Aurora-A to mitotic spindles. *Journal of cell science*, 124, pp.394–404.

Magalska, A. et al., 2014. RuvB-like ATPases Function in Chromatin Decondensation at the End of Mitosis. *Developmental Cell*, 31(3), pp.305–318.

Mejlvang, J. et al., 2014. New histone supply regulates replication fork speed and PCNA unloading. *Journal of Cell Biology*, 204(1), pp.29–43.

Milutinovic, S., Zhuang, Q. & Szyf, M., 2002. Proliferating cell nuclear antigen associates with histone deacetylase activity, integrating DNA replication and chromatin modification. *The Journal of biological chemistry*, 277(23), pp.20974–8.

Minocherhomji, S. et al., 2015. Replication stress activates DNA repair synthesis in mitosis. *Nature*, 528(7581), pp.286–90.

Miotto, B., Ji, Z. & Struhl, K., 2016. Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. *Proceedings of the National Academy of Sciences*, 113(33), pp.E4810–E4819.

Miron, K. et al., 2015. Oncogenes create a unique landscape of fragile sites. *Nature communications*, 6, p.7094.

Misteli, T. & Soutoglou, E., 2009. The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nature reviews. Molecular cell biology*, 10, pp.243–254.

Miuma, S. et al., 2013. Fhit deficiency-induced global genome instability promotes mutation and clonal expansion. *PLoS ONE*, 8(11).

Mohaghegh, P. et al., 2001. The Bloom's and Werner's syndrome proteins are DNA structure-specific helicases. *Nucleic Acids Res*, 29, pp.2843–2849.

Moriarty, H.T. & Webster, L.R., 2003. Fragile sites and bladder cancer. *Cancer Genetics and Cytogenetics*, 140(2), pp.89–98.

Mrasek, K. et al., 2010. Global screening and extended nomenclature for 230 aphidicolin-inducible fragile sites, including 61 yet unreported ones. *Int J Oncol*, 36, pp.929–940.

Murayama, Y. & Uhlmann, F., 2014. Biochemical reconstitution of topological DNA binding by the cohesin ring. *Nature*, 505(7483), pp.367–71.

Musich, P.R. & Zou, Y., 2011. DNA-damage accumulation and replicative arrest in Hutchinson–Gilford progeria syndrome. *Biochemical Society Transactions*, 39(6), pp.1764–1769.

Naim, V. et al., 2013. ERCC1 and MUS81-EME1 promote sister chromatid separation by processing late replication intermediates at common fragile sites during mitosis. *Nat Cell Biol*, 15, pp.1008–1015.

Naughton, C. et al., 2010. Analysis of active and inactive X chromosome architecture reveals the independent organization of 30 nm and large-scale chromatin structures. *Mol Cell*, 40, pp.397–409.

Naughton, C. et al., 2013. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol*, 20, pp.387–395.

Naumova, N. et al., 2013. Organization of the mitotic chromosome. *Science*, 342, pp.948–953.

Neumann, H. et al., 2009. A Method for Genetically Installing Site-Specific Acetylation in Recombinant Histones Defines the Effects of H3 K56 Acetylation. *Molecular Cell*, 36(1), pp.153–163.

Nora, E.P. et al., 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485, pp.381–385.

Ono, T., 2004. Spatial and Temporal Regulation of Condensins I and II in Mitotic Chromosome Assembly in Human Cells. *Molecular Biology of the Cell*, 15(7), pp.3296–3308.

Ono, T., Yamashita, D. & Hirano, T., 2013. Condensin II initiates sister chromatid resolution during S phase. *J Cell Biol*, 200, pp.429–441.

Ozeri-Galai, E. et al., 2011. Failure of origin activation in response to fork stalling leads to chromosomal instability at fragile sites. *Mol Cell*, 43, pp.122–131.

Palakodeti, A. et al., 2009. Impaired replication dynamics at the FRA3B common fragile site. *Human Molecular Genetics*, 19(1), pp.99–110.

Palumbo, E. et al., 2010. Replication dynamics at common fragile site FRA6E. *Chromosoma*, 119(6), pp.575–587.

Panchenko, T. et al., 2011. Replacement of histone H3 with CENP-A directs global nucleosome array condensation and loosening of nucleosome superhelical termini. *Proceedings of the National Academy of Sciences*, 108, pp.16588–16593.

Parada, L.A., McQueen, P.G. & Misteli, T., 2004. Tissue-specific spatial organization of genomes. *Genome biology*, 5(7), p.R44.

Parelho, V. et al., 2008. Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell*, 132(3), pp.422–433.

Pedrali-Noy, G. et al., 1980. Synchronization of HeLa cell cultures by inhibition of DNA polymerase alpha with aphidicolin. *Nucleic acids research*, 8(2), pp.377–387.

Petermann, E. et al., 2010. Hydroxyurea-Stalled Replication Forks Become Progressively Inactivated and Require Two Different RAD51-Mediated Pathways for Restart and Repair. *Molecular Cell*, 37(4), pp.492–502.

Peters, A.H.F.M. et al., 2001. Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell*, 107(3), pp.323–337.

Petryk, N. et al., 2015. Replication landscape of the human genome. *Nature communications*, Under, p.Revision.

Pirzio, L.M. et al., 2008. Werner syndrome helicase activity is essential in maintaining fragile site stability. *J Cell Biol*, 180, pp.305–314.

Polo, S.E. et al., 2010. Regulation of DNA-damage responses and cell-cycle progression by the chromatin remodelling factor CHD4. *The EMBO Journal*, 29(18), pp.3130–3139.

Popuri, V. et al., 2012. Recruitment and retention dynamics of RECQL5 at DNA double strand break sites. *DNA Repair*, 11(7), pp.624–635.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp.841–842.

Ragland, R.L. et al., 2008. Stably transfected common fragile site sequences exhibit instability at ectopic sites. *Genes Chromosomes and Cancer*, 47(10), pp.860–872.

Ran, F.A. et al., 2013. Genome engineering using the CRISPR-Cas9 system. *Nature protocols*, 8(11), pp.2281–308.

Rea, S. et al., 2000. Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*, 406(6796), pp.593–599.

Reijns, M.A.M. et al., 2015. Lagging-strand replication shapes the mutational landscape of the genome. *Nature*, 518(7540), pp.502–506.

Rhind, N. & Gilbert, D.M., 2013. DNA Replication Timing. *Cold Spring Harb Perspect Med*, 3, pp.1–26.

Ricci, M.A. et al., 2015. Chromatin Fibers Are Formed by Heterogeneous Groups of Nucleosomes In Vivo. *Cell*, 160(6), pp.1145–1158.

Rinn, J.L. et al., 2007. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell*, 129(7), pp.1311–1323.

Rogakou, E.P. et al., 1999. Megabase chromatin domains involved in DNA double-strand breaks in vivo. *Journal of Cell Biology*, 146(5), pp.905–915.

Romig, H. et al., 1992. Characterization of SAF-A, a novel nuclear DNA binding protein from HeLa cells with high affinity for nuclear matrix/scaffold attachment DNA elements. *The EMBO journal*, 11(9), pp.3431–3440.

Roukos, V. et al., 2013. Spatial dynamics of chromosome translocations in living cells. *Science (New York, N.Y.)*, 341(6146), pp.660–4.

Roukos, V., Burgess, R.C. & Misteli, T., 2014. Generation of cell-based systems to visualize chromosome damage and translocations in living cells. *Nature protocols*, 9(10), pp.2476–92.

Ruiz-Herrera, A. et al., 2002. Fragile sites in human and Macaca fascicularis chromosomes are breakpoints in chromosome evolution. In *Chromosome Research*. pp. 33–44.

Ryba, T. et al., 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Research*, 20(6), pp.761–770.

Ryba, T. et al., 2011. Genome-scale analysis of replication timing: from bench to bioinformatics. *Nat Protoc*, 6, pp.870–895.

Salic, A. & Mitchison, T.J., 2008. A chemical method for fast and sensitive detection of DNA synthesis in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 105(7), pp.2415–20.

Sander, J.D. & Joung, J.K., 2014. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature biotechnology*, 32(4), pp.347–55.

Saponaro, M. et al., 2014. RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress. *Cell*, 157, pp.1037–1049.

Sengupta, S. et al., 2003. BLM helicase-dependent transport of p53 to sites of stalled DNA replication forks modulates homologous recombination. *EMBO J*, 22, pp.1210–1222.

Sfeir, A. et al., 2009. Mammalian telomeres resemble fragile sites and require TRF1 for efficient replication. *Cell*, 138, pp.90–103.

Shachar, S. et al., 2015. Identification of Gene Positioning Factors Using High-Throughput Imaging Mapping. *Cell*, 162(4), pp.911–923.

Shanbhag, N.M. et al., 2010. ATM-Dependent chromatin changes silence transcription in cis to dna double-strand breaks. *Cell*, 141(6), pp.970–981.

Shintomi, K. & Hirano, T., 2011. The relative ratio of condensin I to II determines chromosome shapes. *Genes & Development*, 25(14), pp.1464–1469.

Shiraishi, T. et al., 2001. Sequence conservation at human and mouse orthologous common fragile regions, FRA3B/FHIT and Fra14A2/Fhit. *Proc Natl Acad Sci U S A*, 98, pp.5722–5727.

Shopland, L.S. et al., 2003. Clustering of multiple specific genes and gene-rich R-bands around SC-35 domains: Evidence for local euchromatic neighborhoods. *Journal of Cell Biology*, 162(6), pp.981–990.

Shreeram, S. et al., 2002. Cell type-specific responses of human cells to inhibition of replication licensing. *Oncogene*, 21(43), pp.6624–6632.

Sirbu, B.M., Couch, F.B. & Cortez, D., 2012. Monitoring the spatiotemporal dynamics of proteins at replication forks and in assembled chromatin using isolation of proteins on nascent DNA. *Nat Protoc*, 7, pp.594–605.

Solomon, D. a et al., 2011. Mutational inactivation of STAG2 causes aneuploidy in human cancer. *Science (New York, N.Y.)*, 333(6045), pp.1039–1043.

Soutoglou, E. & Misteli, T., 2008. Activation of the cellular DNA damage response in the absence of DNA lesions. *Science (New York, N.Y.)*, 320, pp.1507–1510.

Srinivasan, S. V. et al., 2013. Cdc45 Is a Critical Effector of Myc-Dependent DNA Replication Stress. *Cell Reports*, 3(5), pp.1629–1639.

Stanlie, A. et al., 2010. Histone3 lysine4 trimethylation regulated by the facilitates chromatin transcription complex is critical for DNA cleavage in class switch recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 107(51), pp.22190–22195.

van Steensel, B., 2011. Chromatin: constructing the big picture. *The EMBO journal*, 30(10), pp.1885–1895.

Suto, R.K. et al., 2000. Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. *Nature structural biology*, 7(12), pp.1121–1124.

Le Tallec, B. et al., 2013. Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. *Cell Rep*, 4, pp.420–428.

Le Tallec, B. et al., 2011. Molecular profiling of common fragile sites in human fibroblasts. *Nature Structural & Molecular Biology*, 18(12), pp.1421–1423.

Tasara, T. et al., 2003. Incorporation of reporter molecule-labeled nucleotides by DNA polymerases. II. High-density labeling of natural DNA. *Nucleic Acids Research*, 31(10), pp.2636–2646.

Therizols, P. et al., 2014. Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science*, 346(6214), pp.1238–1242.

Tóth, K.F. et al., 2004. Trichostatin A-induced histone acetylation causes decondensation of interphase chromatin. *Journal of cell science*, 117(Pt 18), pp.4277–4287.

Trapnell, C. et al., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3), pp.562–78.

Traverso, G. et al., 2003. Hyper-recombination and genetic instability in BLM-deficient epithelial cells. *Cancer Res*, 63, pp.8578–8581.

Tuduri, S. et al., 2009. Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. *Nat Cell Biol*, 11, pp.1315–1324.

Uemura, T. et al., 1987. DNA topoisomerase II is required for condensation and separation of mitotic chromosomes in S. pombe. *Cell*, 50(6), pp.917–925.

Ui, A., Nagaura, Y. & Yasui, A., 2015. Transcriptional Elongation Factor ENL Phosphorylated by ATM Recruits Polycomb and Switches Off Transcription for DSB Repair. *Molecular Cell*, 58(3), pp.468–482.

Untergasser, A. et al., 2012. Primer3-new capabilities and interfaces. *Nucleic Acids Research*, 40(15).

Ushiki, T. & Hoshi, O., 2008. Atomic force microscopy for imaging human metaphase chromosomes. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 16(3), pp.383–96.

Vannier, J.B. et al., 2012. RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity. *Cell*, 149, pp.795–806.

Venkatraman, E.S. & Olshen, A.B., 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6), pp.657–663.

Walter, J. & Newport, J., 2000. Initiation of eukaryotic DNA replication: origin unwinding and sequential chromatin association of Cdc45, RPA, and DNA polymerase alpha. *Mol Cell*, 5, pp.617–627.

Wang, L. et al., 1999. Allele-specific late replication and fragility of the most active common fragile site, FRA3B. *Hum Mol Genet*, 8, pp.431–437.

Wechsler, T., Newman, S. & West, S.C., 2011. Aberrant chromosome morphology in human cells defective for Holliday junction resolution. *Nature*, 471, pp.642–646.

Wei, Y. et al., 1999. Phosphorylation of histone H3 is required for proper chromosome condensation and segregation. *Cell*, 97, pp.99–109.

West, M.H. & Bonner, W.M., 1980. Histone 2A, a heteromorphous family of eight protein species. *Biochemistry*, 19(14), pp.3238–45.

Williamson, I. et al., 2012. Anterior-posterior differences in HoxD chromatin topology in limb development. *Development*, 139(17), pp.3157–3167.

Williamson, I. et al., 2014. Spatial genome organization: Contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes and Development*, 28(24), pp.2778–2791.

Wirbelauer, C., Bell, O. & Schübeler, D., 2005. Variant histone H3.3 is deposited at sites of nucleosomal displacement throughout transcribed genes while active histone modifications show a promoter-proximal bias. *Genes and Development*, 19(15), pp.1761–1766.

Xu, Y. et al., 2012. Histone H2A.Z Controls a Critical Chromatin Remodeling Step Required for DNA Double-Strand Break Repair. *Molecular Cell*, 48(5), pp.723–733.

Yu, A. et al., 2000. Activation of p53 or loss of the Cockayne syndrome group B repair protein causes metaphase fragility of human U1, U2, and 5S genes. *Molecular Cell*, 5(5), pp.801–810.

Zhang, Y. et al., 2012. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, 148(5), pp.908–921.

Ziv, Y. et al., 2006. Chromatin relaxation in response to DNA double-strand breaks is modulated by a novel ATM- and KAP-1 dependent pathway. *Nature cell biology*, 8(8), pp.870–876.